THEORETICAL POPULATION AND QUANTITATIVE GENETICS AND ANIMAL IMPROVEMENT

by

WILLIAM GEORGE HILL

B.Sc.(London)
M.S. (California)
Ph.D.(Edinburgh)

Thesis presented for degree of Doctor of Science

University of Edinburgh 1976

CONTENTS

ABSTRACT

The thesis comprises a collection of thirty research papers divided into four groups, which cover a range of topics from linkage disequilibrium due to genetic drift, through long term selection for quantitative traits to aspects of the design of applied breeding programmes. All the papers have a mathematical or statistical content and do not include experimental results. Much of the analysis is of problems in finite population theory and demonstrates the important effect that small population number can have on the rates, variability and limits of genetic change.

In group I, entitled "Long term selection for quantitative traits in finite populations" there are a series of papers on the rates and limits of response to selection, in which there is a detailed analysis of the changes in gene frequency at one or a few loci which contribute a small part of the total genetic variation. Simple approximations using selective values are shown to adequately describe truncation selection in finite populations. The effects of different degrees of dominance on long term changes in gene frequency and population mean are described, and the efficiency of alternative methods of improvement to selection in a single population, such as selection on crossbred performance or subdivision of the population into small lines, are compared. Tight linkage is shown to reduce selection limits for additive genes initially in linkage equilibrium.

Group II, "Design and analysis of experiments to estimate genetic parameters" includes two papers on analysis of data obtained from a single set of parents and progeny in which it is shown that sampling errors of heritability estimates can be reduced by selection among the parents, or by combining regression and sib covariance

estimates. The main theme of the group is, however, the design and analysis of selection experiments for quantitative traits. Formulae are developed for predicting the variance due to genetic drift between conceptual replicate populations, so that the standard error can be computed of estimates of heritability or other parameters obtained from the results of the experiments. Conventional analyses are shown to considerably underestimate the precision of estimates, and unbiassed methods are suggested. These results, together with a formula developed for the effective population size in populations with overlapping generations, are brought together in a review paper on estimating genetic change.

Some applied papers on animal improvement are in group III "Topics in the design of breeding programmes". They include an analysis of the discounting method of financial evaluation of breeding programmes with examples of its use in continuous breeding programmes and an extension to breed comparisons. The possible role of new synthetic breeds or populations and the possible use which could be made of the technique of superovulation and ovum transplantation in animal improvement programmes are discussed. There are also more basic papers in the group. A new matrix method is developed for describing the rate and pattern of response to selection with overlapping generations, and there are analyses of the effects of sampling errors of parameter estimates on the efficiency of selection indices to improve a single trait using information on relatives or on secondary traits.

Group IV, "Linkage disequilibrium: generation by genetic drift and statistical tests", mainly comprises a series of papers which predict the amount of disequilibrium, or non-random

association, between genes having no selective effect which is caused by sampling in populations of finite size. Starting from a basic model of two linked loci each with two alleles, the results are extended to three or more loci and to two loci with a conceptually infinite number of alleles. In populations initially in linkage equilibrium drift causes no mean change, but a variance in disequilibrium is induced by drift. It is shown that, within populations segregating at the two loci, the mean squared correlation of gene frequencies between the loci asymptotes at approximately 1/(4 x population size and recombination fraction) and this result can be extended to more complex situations. Two papers are included which give methods of estimating and testing for disequilibrium in diploid populations, and it is demonstrated that estimation from diploids can be equally efficient, per observation, as the use of extracted chromosomes.

INDEX OF PAPERS

GROUP I : LONG TERM SELECTION FOR QUANTITATIVE TRAITS IN FINITE POPULATIONS.

GROUP II : DESIGN AND ANALYSIS OF EXPERIMENTS TO ESTIMATE GENETIC PARAMETERS.

10. HILL, W.G. 1971. Design and efficiency of selection experiments for estimating genetic parameters. Biometrics 27: 293-311.

11. HILL, W.G. 1972. Estimation of realised heritabilities from selection experiments. I. Divergent selection. Biometrics 28: 747-765.

12. HILL, W.G. 1972. Estimation of realised heritabilities from selection experiments. II. Selection in one direction. Biometrics 28: 767-780.

13. HILL, W.G. 1974. Variability of response to selection in genetic experiments. Biometrics 30: 363-366.

14. HILL, W.G. 1972. Effective size of populations with overlapping generations. Theor.Pop.Biol. 3: 278-289.

15. HILL, W.G. 1972. Estimation of genetic change. I. General theory and design of control populations. Anim.Breed.Abstr. 40: 1-15.

GROUP III : TOPICS IN THE DESIGN OF BREEDING PROGRAMMES

16. HILL, W.G. 1971. Investment appraisal for national breeding programmes. Anim.Prod. 13: 37-50.

17. HILL, W.G. 1974. Prediction and evaluation of response to selection with overlapping generations. Anim.Prod. 18: 117-139.

18. HILL, W.G. 1971. Theoretical aspects of crossbreeding. Ann.Genet. Sel.Anim. 3: 23-34.

19. HILL, W.G. 1974. Size of experiments for breed or strain comparisons. Proc.Symp.Breed Evaluation and Crossing Expts.with Farm Animals, pp.43-54.

20. LAND, R.B. and HILL, W.G. 1975. The possible use of superovulation and embryo transfer in cattle to increase response to selection. Anim.Prod. 21: 1-12.

21. SALES, J. and HILL, W.G. 1976. Effect of sampling errors on efficiency of selection indices. I. Use of information from relatives for single trait improvement. Anim.Prod. 22: 1-17.

22. SALES, J. and HILL, W.G. 1976. Effect of sampling errors on efficiency of selection indices. II. Use of information on associated traits for improvement of a single important trait. Anim.Prod. 23 (in press).

GROUP IV: LINKAGE DISEQUILIBRIUM : GENERATION BY GENETIC DRIFT AND STATISTICAL TESTS

23. HILL, W.G. and ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. Theor.Appl.Genet. 38: 226-231.

24. HILL, W.G. 1969. Maintenance of segregation at linked genes in finite populations. Jap.J.Genet. 44: Suppl.1: 144-151.

25. HILL, W.G. 1974. Disequilibrium among several linked neutral genes in finite population. I. Mean changes in disequilibrium. Theor.Pop.Biol. 5: 366-392.

26. HILL, W.G. 1974. Disequilibrium among several linked neutral genes in finite population II. Variances and covariances of disequilibria. Theor.Pop.Biol. 6: 184-198.

27. HILL, W.G. 1976. Non-random association of neutral linked genes in finite populations. In Population Genetics and Ecology, ed. S. Karlin and E. Nevo. Academic Press, New York, pp.339-376.

28. HILL, W.G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Pop.Biol. 8: 117-126.

29. HILL, W.G. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity 33: 229-239.

30. HILL, W.G. 1976. Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 31:881-888.

REVIEW OF PAPERS

The majority of papers in this collection are concerned with problems of finite population size, and demonstrate the influence that small population number can have on rates, variability and limits of genetic change. Such studies were largely motivated by the fact that most experimental and commercial populations at the nucleus level comprise a limited number of breeding individuals. All the papers included are theoretical, in that they are mathematical or statistical rather than experimental, but range from basic population genetics to applied animal improvement. The papers have been classified into four groups, but these are not mutually exclusive and several papers could have been put into more than one group.

Group I. Long term selection for quantitative traits in finite populations.

In this group of papers the effects of population number on the rates and limits of response to selection for quantitative traits are discussed. The other aspects of the models studied include degree of dominance, linkage and the structure of the breeding programme. In order to obtain insight into the processes involved, the papers comprise detailed analyses of the changes in gene frequency at one or a few loci which contribute only a small part of the total genetic variation, rather than attempts to simulate whole model organisms. From a mathematical viewpoint, they are all studies in applied stochastic processes, and illustrate the use of transition probability matrices and Monte Carlo methods in the analysis of Markov chains.

Paper 1 comprises a study of the effects of linkage on rates and limits to artificial selection in finite populations, using

a model of just two loci with additive genes. With genes initially in linkage equilibrium, it was found that tight linkage would be expected to reduce total progress, although the early rate of response would be little altered. Whilst the influence of linkage or change of frequency would be most marked for genes having small effect on the trait under selection but which were linked to another of large effect, the influence on the mean of the trait would be greatest when genes were of more nearly equal effect at low initial frequency. Whilst the model discussed in paper 1 is more complicated than some in the following papers, it was written earlier and includes results used subsequently (e.g. papers 3, 23).

A basic theory is given in paper 2 of the change in gene frequency expected at a single locus when truncation selection is practised in a finite population. For one cycle of selection the probability distribution is computed for the number of individuals of each genotype which are selected, and the mean gene frequency in the progeny generation thereby obtained. By incorporating the sampling distribution of progeny from parents, transition probabilities for the numbers of each genotype in successive generations are derived. Numerical checks showed that formulae both for selective values derived using infinite population theory and probabilities of fixation based on fitness rather than artificial selection models gave very good approximations for truncation selection in small populations. This was reassuring as these approximations had already been widely used.

The rates of change in the frequency of non-additive genes at single loci together with the corresponding changes in their contribution to mean performance, which is not necessarily in the same or in uniform direction, are discussed for finite populations

in papers 3 and 4. The dependence of the half-life of response, the time taken to get half way to the limit, on the intensity of selection, population size, initial frequency and direction of dominance is illustrated in paper 3. Whilst it should be possible to estimate average gene effects from half-lives, it is clear that many assumptions have to be made. Paper 4 comprises an analysis of the effect of selection in reducing the rate of inbreeding depression for loci at which heterozygotes are superior. The results show that the reduction in rate is due to maintenance of segregation if the equilibrium gene frequency is near one half, but to fixation of the better homozygote when the equilibrium departs far from one half. Numerical results are also included in paper 4 showing that appropriately constructed models based on one sex give good approximations to those with two sexes.

In addition to selection intensity and population size, the breeder is able to alter the population structure. To improve crossbreds, pairs of populations may be selected on cross performance by reciprocal recurrent selection (RRS) or one may be crossed to an inbred tester (RST). The long term consequences of such selection in finite populations are analysed in paper 5 and compared with selection on pure line performance (PLS). General conclusions were difficult to obtain because the relative efficiencies of the alternative population structures depend on the degree of dominance. In terms of the parameter combination, population size $\times$ selection intensity, which in practice is likely to be largest with PLS, it was found that RRS is only slightly more efficient than PLS for completely dominant genes; and with overdominant genes at equilibrium, such as in a plateaud population, RST is predicted to give higher initial rates of response than RRS, but the same limit. In paper 6 the effects are considered of subdividing a population into small lines

to enable selection between them to utilise the variation caused by genetic drift. For models of additive or completely dominant genes it was found that such subdivision and crossing schemes were unlikely to be useful in the long term, except for the more rapid elimination of deleterious recessive genes at low initial frequency. Some short term advances may be gained by using only the best lines, but in the long term more intense selection within populations may be preferable to between-line selection. In both papers 6 and 7 an approximate theory based on degrees of inbreeding was found to give a good impression of the pattern of response.

Paper 7 is included in this group because fixation probabilities are computed for finite populations, but as it involves a model in which population size varies stochastically, is likely to be of more relevance to natural than artificial selection. The work is primarily methodological, but demonstrates that approximations based on the arithmetic mean of selective value and the harmonic mean of population size gives good predictions of fixation probabilities.

Group II. Design and analysis of experiments to estimate genetic parameters.

Whilst response over many generations to a possible limit may ultimately be of some concern to the breeder, he has first to predict responses and try to use an efficient selection scheme in the short term. For predictions over a few generations parameters such as heritabilities, genetic correlations and variances may be sufficient, without trying to consider the effects of individual genes. This group of papers deals essentially with estimation of parameters such as heritability. A theory is developed to predict the effects of genetic drift on the response to continued selection in finite

populations and thus the sampling variances of heritability estimates obtained from selection experiments. Correction for environmental change is necessary if genetic change is to be estimated, so a paper on design of control populations and another on effective population size are included in this group.

The sampling variance of the estimate of heritability from the regression of progeny on parent performance can be reduced for a given total number of records by selecting only the extreme high and low ranking individuals as parents. Paper 8 gives formulae and examples for computing the optimal proportion of potential parents to select and number of progeny to record from each parent. The designs are rather robust to poor a priori predictions of the unknown parameter, and variances of heritability estimates per observation can be roughly halved relative to taking parents at random.

When data are available on parents and progeny, heritability and similar parameters can be estimated both from the covariance among sibs and from the regression of progeny on parent performance, yet this is rarely done in practice. In paper 9 such pairs of estimates are shown to be correlated, so care has to be taken when making genetic inferences about their relative magnitude. Maximum likelihood and simpler weighting procedures are described which make full use of the data, and optimum designs described when maximum likelihood methods are to be adopted.

In papers 10-13 a theory is developed for predicting the variability in response of quantitative traits in selection experiments or breeding programmes that are conducted in small populations, and for analysing the results. Since genetic sampling occurs each generation the variance between actual or conceptual

replicate populations in mean genetic merit increases in successive generations and may be much larger than the variance of estimate of the mean genotype from measurement of phenotype on only a small number of individuals. Furthermore, the genetic drift induces a correlation between performance in different generations.

The main formulae for the error structure of the generation means is derived in paper 10. The results for one generation of selection is obtained by straightforward statistical methods, but several assumptions are made to enable the analysis to be extended to successive generations, the most important being that the change in variance within populations is small and can be ignored. The results therefore do not hold in the long term, any such predictions would require information on gene effects and frequencies. Paper 10 also includes values for the optimal selection intensity for estimating realised heritability, the regression of response on selection differential, and it is demonstrated that selection experiments can be efficient for estimating both heritabilities and genetic correlations.

The theoretical predictions of paper 10 are used in papers 11 and 12 to investigate the efficiency of alternative methods of analysing data from selection experiments for parameter estimation. The selection scheme discussed in paper 11 is of divergent selection, i.e. a pair of lines selected in opposite directions from the same base population; and in paper 12 unidirectional selection is assumed to be practised so, if no control population is maintained, variation due to environment common to all members of a generation may be present. It was found that, because the error structure of generation means is correlated due to genetic drift, alternative linear regression estimates do not differ much in efficiency from each other or from a maximum likelihood estimator. But, more importantly, if the usual linear regression model of independent equal errors is assumed for the generation means the computed sampling

variance for the estimate of, for example, realised heritability is expected to be very much smaller than its real value. Relatively unbiassed methods of computing such sampling variances are suggested.

When estimating parameters from a selection experiment, the variances of response required are those conditional on the selection differentials applied and these conditional variances are used in papers 10-12. The variance in performance observed between small replicate lines is also increased by random variation in selection differential. Thus formulae are developed in paper 13 for the variance-covariance structure of generation means, not conditioned by the selection differential applied. The results do not differ greatly from the corresponding conditional values.

Formulae are derived in paper 14 for the effective population size of random mating populations in which generations overlap. With a few limiting assumptions, it is shown that the effective population number with overlapping generations is the same as that of a population with discrete generations having the same number of individuals entering each generation and the same variance of lifetime family size. Even when there are no real differences between individuals in viability or fertility, due to chance differences in age at death the variance in lifetime family size is likely to be larger in populations with overlapping generations. The results for effective size hold only asymptotically after many generations, and those for early generations are still to be derived.

Although not relevant to the main theme of this group, paper 14 is included here because its results are used in paper 15, in which problems of estimating genetic change, specifically the removal of environmental trend, and the design of control populations are discussed. This is primarily a review paper, but includes new material on the variation between the means of selected and control populations,

utilising the results of paper 10, and on the effects of non-random mating and enforced zero selection differentials on drift variances in control populations.

Group III. Topics in the design of breeding programmes

The papers of group III cover a variety of problems, but all are concerned directly or indirectly with optimal design of breeding programmes. There is an analysis of evaluation of financial returns from breeding programmes with examples taken from continuous selection schemes and breed comparisons. The possible role of synthetic populations and of breeding programmes to utilize new techniques in reproductive physiology are discussed. There are also more basic papers on the rate and pattern of response expected from selection in populations with overlapping generations, and on the effects of sampling errors in parameter estimation on the efficiency of selection indices.

Alternative breeding programmes may differ substantially in costs, therefore the criterion for designing one to use in practice can not simply be to maximise the rate of genetic progress. Paper 16 includes an outline for animal breeders of the use of the method from management accounting of discounting monetary returns expected in future years back to present value, so that all costs and returns can be combined at the same base points. Two contrasting examples from possible genetic improvement programmes for beef production are used for illustration. The analysis shows clearly the important contribution made by early returns and the long time before some schemes break even. Although the effect of some changes in assumptions were examined in paper 16, showing one scheme likely to be better than the other over a wide range of assumptions, this is not always the case with such analyses. The discounting method is not a panacea: some assumptions, such as of market size are very

critical, and there are alternative ways of assessing the results, e.g. internal yield and net return at fixed discount rate. With so much emphasis put by discounting on short term response, it becomes immaterial in the calculations whether the population runs out of variation after twenty or thirty years; posterity seems to get ignored. Whilst this author has concentrated more on short term problems after writing paper 16, there still remains a need to balance short term and long term responses.

In populations in which generations overlap, selection practised among the current crop of young animals does not produce a uniform improvement of all animals of the next crop, because these have a range of parental age. Response in early generations of a programme is therefore erratic, as illustrated in paper 16, and if returns are heavily discounted this could affect economic assessment of a programme. Formulae for the asymptotic rate of response in a continuing programme are well known, and several informal methods have previously been described to predict the pattern of response in populations having overlapping generations, but in paper 17 a general and formal method is derived. It is based on a modification of Leslie matrices, which are widely used in population dynamics, to show the contribution of genes between individuals of different ages in successive breeding seasons. These matrices, previously used during the computation of effective population size in paper 14, display the breeding structure; and standard matrix operations enable predictions of responses to selection among individuals born in each successive breeding season and of the lag in response between animals of different age and level in the breeding and multiplication pyramid.

Paper 18 comprises a review and discussion of various aspects of crossbreeding, but includes some new material on the

possible improvement of performance using novel synthetic breeds or populations. It is argued firstly that, although synthetics might show more variability than their parents, this would be difficult to prove in experiments of feasible scale, and secondly that any period of reduced selection while the synthetic was formed, particularly if its performance was behind that of the best available commercial population, would take many years to make up. During this time the synthetic would make no commercial contribution.

Many breed and strain evaluation trials are being designed and carried out, but the justification for the size of individual experiments is not always clear. In paper 19 an attempt is made to determine the optimal number of animals to include in an experiment by balancing costs and the expected financial returns to be made when decisions are taken on breed replacement as a result of the trial. Many assumptions have to be made, but some of the results are quite robust. As with the arguments of paper 16, it is perhaps more important for the breeder to put the design problems into a financial perspective rather than be concerned with details.

The techniques of superovulation and ovum transplantation offer the potential to breed many more progeny from selected females than with natural ovulation, and their possible value in continuing cattle improvement programmes is discussed in paper 20. In a scheme considered to be feasible, rates of genetic progress for traits of the growing animal could be nearly doubled by use of ovum transfer from individuals selected on a performance test. By recording successive ovulations by laparoscopy and selecting the best females for induced superovulation and transfer, rates of response for twinning frequency might be greatly increased over that from conventional schemes, but even so the rates of response may be too low to be of value.

The remaining papers of the group, papers 21 and 22, report analyses of the effects of sampling errors of genetic parameter estimates on the efficiency of selection indices used to combine information on relatives and/or several traits when ranking individuals for selection. Three quantities are compared: the response possible with the optimum index, i.e. an index computed using the parameter values, the response predicted from using the index computed from parameter estimates, and the response likely to be achieved when the latter index is used in the population. The combination of data on a single trait on an individual and its sibs is discussed in paper 21, where it is found that, although predictions of absolute response are much affected by errors in estimates, the prediction of response from selection on the index relative to that on individual performance is less sensitive and, more importantly, the response achieved when the index is used is very robust to sampling errors in the estimates of parameters. In contrast, where the index is used to augment information from a trait of economic importance by that from a second of no direct importance, but perhaps correlated with the first, the analysis in paper 22 demonstrates that substantial reductions may be obtained in the response achieved if there are imprecise parameter estimates. If the traits are really uncorrelated, any error in the estimate of correlation will show the predicted index to be more efficient than selection on a single trait, when in fact it is not. It turns out that the expected benefit from inclusion of the second trait then equals its real loss. Studies on other examples of index use are in progress.

Group IV. Linkage disequilibrium: generation by genetic drift and statistical tests

Most of this final group of papers comprise a study of

the amount of linkage disequilibrium, i.e. non-random association, between loci likely to be caused by sampling in populations of finite size. With minor exception, no selection is included in the models. These papers are motivated by the need for an adequate population genetic theory of neutral genes in finite populations against which evidence can be tested. The analysis uses recurrence relations among moments of the gene and chromosome frequency distributions in successive generations, except where moments are required only among those populations still segregating at the relevant loci, when Monte Carlo methods are used. Two papers are also included on statistical methods for estimating the degree of linkage disequilibrium and testing for its presence, the choice of quantities to describe finite population predictions being influenced by the statistical tests available for analysing data.

Paper 23 includes one of the first published demonstrations that linkage disequilibrium (D) could be generated between neutral genes by genetic drift. (It is being reprinted in a volume of collected papers "Stochastic models in population genetics", edited by W.-H. Li). Although the mean value of D remains zero over replicate populations, its variance does not. A moment generating matrix is used for computing the mean of D^2 and explicit results are given in a special case. The squared correlation of gene frequencies (r^2) in populations segregating at both loci is adopted as a statistic less dependent on absolute gene frequency than D^2. Although Monte Carlo simulation was required to find the mean of r^2 each generation, a useful simple approximation to its asymptotic value, 1/(4 x population size x recombination fraction) is given. This paper demonstrates that the presence of linkage disequilibrium could not necessarily be attributed to selection.

Whilst paper 23 includes some discussion of the effects of selection for heterozygotes, where it shows for a range of parameter values that mean values of r^2 are not greatly different from that expected for neutral genes, the point is made in paper 24 that tight linkage between loci each with heterozygote superiority retards fixation in small populations. Further work on the joint effects of selection and drift are planned.

The moment generation matrix approached is used in two essentially mathematical papers, papers 25 and 26, to extend to more than two loci the study of means, variances and covariances between neutral linked genes in finite populations. As with two loci, formulae for the moments depend only on products of population size and map length. For more than three loci, expressions for the mean multi-locus disequilibrium are shown in paper 25 to involve products of disequilibria among fewer loci. Therefore the rate of disappearance of multi-locus disequilibrium from populations, such as crosses, not initially in equilibrium is at least as fast as the sum of the rates of loss of disequilibrium between constituent pairs of loci. With equally spaced loci the rate is roughly proportional to the number of loci defined in the multi-locus disequilibrium. For populations initially in equilibrium it is shown in paper 26 that the covariance between disequilibria at different pairs of loci remains zero. The variance of the three locus disequilibrium reaches a maximum in earlier generations than does that for two loci, and the value and time at this maximum depends mostly on the map length between the more distant loci.

In paper 25 and 26 moments of disequilibria are computed over all populations, whether segregating or not, but in paper 27 the analysis of the moments in segregating populations alone is given in

detail for three and in principle for more neutral loci. Although analytically more difficult to obtain, results for segregating population are likely to be of more practical use since it is in this subset that the experimentalist tests for disequilibrium. With three loci having two alleles at each, the chromosome frequencies form a 2 x 2 x 2 contingency table, and since there are alternative ways recognised for statistically analysing such data to test for the three-way association, a part of paper 27 is devoted to considering how best multi-locus results should be presented and interpreted. Some simple examples show that the quantitative measure of three-way association with vey high linkage depends on the methods of analysis; the method adopted is that most acceptable statistically, but less tractable computationally. The Monte Carlo results show that with tight linkage among neutral genes in segregating finite populations, the three locus associations makes a very small contribution to the total contingency chi-square for lack of goodness-of-fit to independence of gene frequencies.

In the models of the preceding papers of this group only two alleles are present at each locus. The analysis given in paper 28 is essentially a two-locus extension of studies in which each mutational event produces an allele not currently present in the population, and all these alleles, conceptually infinite in number, have no effect on fitness. Using moment generating matrices, simple relationships are obtained between the expected sum of squares of disequilibria among all possible pairs of alleles and the products of the heterozygosities at the two loci. Unless population size and recombination fraction are both very small, the ratio of the expectations of disequilibria and products of heterozygosities approximates $1/(4 \times \text{population size} \times \text{recombination fraction})$, just as in the two allele model. Further

unpublished work has shown that this is also the approximate asymptotic value within segregating populations of the expectation of the ratio of disequilibria and heterozygosity, and of the contingency table chi-square, standardised by degrees of freedom, in a test for random association of alleles.

Methods of statistical analysis of data on genotype frequencies at two or more loci obtained from, for example, gel electrophoresis of enzyme polymorphisms are given in the remaining papers. These show how linkage disequilibrium (D) can be estimated and its presence tested. In paper 29 maximum likelihood procedures are outlined for estimating D from observations on diploid individuals at two loci each with two alleles which are codominant and/or dominant. The sampling variances are compared for estimates of D from diploid data or from extracted chromosomes, a feasible technique in Drosophila. When the loci are actually in equilibrium it turns out that the relative efficiencies of the two methods are the same per individual typed, but the laboratory work using diploids is, of course, much less per individual observation. A general scheme is given in paper 30 for analysis of data on diploids to estimate chromosome frequencies and test for linkage disequilibrium between any number of loci with two or more codominant alleles, although the details are given only for three loci each with two alleles. Estimation is by maximum likelihood and likelihood ratio tests are used to distinguish between different assumptions of dependence of frequencies at the constituent loci. The analysis extends to diploids the methods of multi-dimensional contingency tables, and some justification for the procedure is given in paper 27.

STATEMENT OF AUTHORSHIP

In those papers of which I was the sole author the work was initiated, carried out and written up by myself, apart from technical help. It is not, of course, possible to reliably attribute contributions to separate authors in the studies reported under joint authorship because in each case I was working closely with my co-author and the process of development was interactive. I hope my recollection of individual responsibility is fair. Papers 1, 4 and 23 were written with Alan Robertson, who was my Ph.D. supervisor for the work reported in paper 1. In each case he was mainly responsible for initiating the study, but most of the results were obtained by me. Pape 6 with Fernando Madalena is based on part of his Ph.D. study; the work was initiated jointly, largely executed by him and the paper was written up by myself. Paper 6 also includes a section by Alan Robertson, which is acknowledged therein. I was responsible for initiating paper 9 and obtained most of the basic results, but the section on design is due to Frank Nicholas. Paper 20 was initiated and carried out jointly with Roger Land, my particular responsibility being the calculations. The work of papers 21 and 22 with Jill Sales was largely initiated and written up by me and I also derived a few of the results.

The work included has not been submitted for other degrees, with the following exceptions. Paper 1 is based on part of my Ph.D. thesis (University of Edinburgh, 1965), paper 6 is based on part of the Ph.D. thesis of Madalena (University of Edinburgh, 1970) and paper 9 is included as an appendix to the Ph.D. thesis of Nicholas (University of Edinburgh, 1974).

I have had the benefit over the years of innumerable stimulating discussions with colleagues and visitors at the Institute of Animal Genetics, our neighbours at A.B.R.O. and members of the Statistical Laboratory in Ames. For their counsel, I wish to thank Alan Robertsson, in particular, and Douglas Falconer, Oscar Kempthorne and Roger Land among many others. I have had much highly competent technical assistance from Jenny Smith, Marjorie McEwan and Kathy Burgoyne, and all the graphs have been excellently drawn by E.D. Roberts (Robby). Finally I would like to thank my family for allowing me time to work.

WILLIAM G. HILL May 1, 197

1

The effect of linkage on limits to artificial selection

by

William G. Hill and Alan Robertson

Genet. Res., Camb. (1966), **8**, *pp.* 269–294
With 11 *text-figures*
Printed in Great Britain

The effect of linkage on limits to artificial selection

By W. G. HILL and ALAN ROBERTSON*
Institute of Animal Genetics, Edinburgh, 9

(*Received* 1 *April* 1966)

1. INTRODUCTION

A theory of limits to artificial selection in small populations was given by Robertson (1960) in terms of single genes, and was extended to selection for a quantitative character governed by many loci by ignoring linkage and epistatic interactions between loci. In this paper we include the effect of linkage in a very simple situation, that of two additive loci, though it is hoped to deal with more complex models in further papers. No general algebraic solution to this problem has been found, so that most of our information has come from Monte Carlo simulation on computers. When there is no recombination between the two loci, an algebraic treatment has been developed which will be described in a later paper.

Griffing (1960) investigated the effect of linkage on response to artificial selection in infinitely large populations, assuming that gene effects were small enough that changes in genetic parameters, other than the population mean, could be ignored. Using a model of two loci in an infinite population, Nei (1963) and Felsenstein (1965) have developed formulae for the effect of directional selection on changes in linkage disequilibrium. But, in infinite populations, linkage cannot affect the selection limit but only the rate of advance to that limit. Simulation by Monte Carlo methods has shown that, though populations may initially be in linkage equilibrium, the advance under selection can be reduced when genes are tightly linked, even with no interactions between loci (Martin & Cockerham, 1960; Qureshi, 1963). These workers used models in which the initial gene frequency of 0·5 and the effect on the character under selection were the same for all loci. Latter (1965*b*), using only two loci, considered the consequences of varying the initial gene frequency though this and the effect on the character under selection were the same for both loci. We shall also restrict ourselves here to two loci with additive action, but shall not restrict the effects of the loci on the character under selection or the initial gene frequency. We will in general assume that the population is initially in linkage equilibrium.

2. BASIC THEORY

To give a framework for the theoretical consideration of the problem with two loci, it will be useful to repeat some of Robertson's earlier conclusions on selection in a finite population at a locus with additive gene action, which relied heavily on a paper by Kimura (1957). The basic concept underlying this is the gene frequency

* Member of the Agricultural Research Council Unit of Animal Genetics.

distribution. This can be regarded as either that of the frequencies at equivalent loci in one population or that at a single locus replicated in many equivalent populations. Similarly, the chance of fixation when the selection limit is reached can be considered either as the proportion of such loci fixed in the same direction in a single line, or as the proportion of replicate lines in which the same allele is fixed. The situation in which no further selection response can be made but in which not all loci are fixed, due to heterozygote superiority or opposing natural selection, will not be discussed.

At a locus at which there is additive action in selective advantage (as would be brought about by artificial selection acting on a locus with additive effect on the character under selection), the change in the distribution (ϕ) of gene frequencies with time can be described reasonably well by the diffusion equation

$$\frac{\partial \phi}{\partial (t/N)} = \frac{1}{4}\frac{\partial^2}{\partial p^2}[p(1-p)\phi] - \frac{Ns}{2}\frac{\partial}{\partial p}[p(1-p)\phi] \tag{1}$$

where p is the gene frequency, t is the time in generations, N is the population size and s is the difference in selective advantage between the two homozygotes. From a given initial gene frequency, the pattern of the selection process is then entirely determined by the parameter Ns on a time scale t/N. Kimura (1957) showed that the chance of eventual fixation, $u(p_0)$, of a gene with initial frequency p_0 is then a function only of Ns and is given explicitly by

$$u(p_0) = \frac{1-e^{-2Nsp_0}}{1-e^{-2Ns}} \tag{2}$$

Examination of equation (1) shows that any computer simulation of the selection process need only be done at one population size as the above generalization allows extrapolation to all values of N and s. In practice this is limited by the restriction that s shall not be greater than unity. From equation (2) it can be shown that if Ns is small ($< 0{\cdot}5$) then the expected change in gene frequency at the limit is $2N$ times the change in gene frequency in the first generation and that the time for the gene frequency to change by half this amount is $1{\cdot}4N$ generations. Under most conditions, this value is an upper limit for the 'half-life' of the selection process.

When more than one locus segregates, the differential equation describing the selection process can be written in terms of gametic frequencies in the general form as follows (e.g. Kimura, 1955):

$$\frac{\partial \phi}{\partial t} = \frac{1}{2}\sum_{j=1}^{n}\frac{\partial^2}{\partial f_j^2}[V(\delta f_j)\phi] + \sum_{j<k}\sum\frac{\partial^2}{\partial f_j\,\partial f_k}[\mathrm{cov}(\delta f_j, \delta f_k)\phi] - \sum_{j=1}^{n}\frac{\partial}{\partial f_j}[M(\delta f_j)\phi] \tag{3}$$

where $\phi(f_1, f_2 \ldots f_n, t)$ is the density function of the distribution of gametic frequencies, f_j, at time t. The dimension (n) of the equation is the number of degrees of freedom amongst the gametic frequencies. Thus for two loci, each with two alleles, $n = 3$. From the multinomial distribution, the variance of change in gametic frequency is given by

$$V(\delta f_j) = [f_j(1-f_j)]/2N$$

and the covariance of changes by

$$\text{cov}\,(\delta f_j, \delta f_k) = -f_j f_k/2N$$

For the simplest model of two loci each with two alleles, let the frequencies of the gametes, AB, Ab, aB and ab be f_1, f_2, f_3 and f_4 respectively. Also let p and q be the frequencies of the alleles A and B, and define linkage disequilibrium by the determinant $D = f_1 f_4 - f_2 f_3$. Finally, assume that these loci have additive selective values r and s, the differences in selective values between the homozygotes at loci A and B respectively, and let c be the recombination fraction between these loci, assumed to be the same for both sexes. Then

$$\begin{aligned} M(\delta f_1) &= \tfrac{1}{2} f_1[r(1-p)+s(1-q)]+\delta D \\ M(\delta f_2) &= \tfrac{1}{2} f_2[r(1-p)-sq]-\delta D \\ M(\delta f_3) &= \tfrac{1}{2} f_3[-rp+s(1-q)]-\delta D \end{aligned} \tag{4}$$

and

$$\delta D = -cD\{1+\tfrac{1}{2}[r(1-2p)+s(1-2q)]\} \tag{5}$$

In equations (4) and (5), r and s are assumed small so that terms in the denominator have been ignored. Also, for the diffusion equation to hold, r, s and c must be small such that terms in their products can be ignored relative to $1/N$. Thus we can take

$$\delta D = -cD$$

Multiplying (3) by N and inserting the above equations, we obtain for two additive loci

$$\begin{aligned} \frac{\partial \phi}{\partial (t/N)} = {} & \frac{1}{4}\sum_{j=1}^{3} \frac{\partial^2}{\partial f_j^2}[f_j(1-f_j)\,\phi] - \frac{1}{2}\sum_{j<k}\sum \frac{\partial^2}{\partial f_j\,\partial f_k}[f_j f_k \phi] \\ & -\tfrac{1}{2}Nr\left\{\frac{\partial}{\partial f_1}[f_1(1-p)\,\phi]+\frac{\partial}{\partial f_2}[f_2(1-p)\,\phi]-\frac{\partial}{\partial f_3}[f_3 p\phi]\right\} \\ & -\tfrac{1}{2}Ns\left\{\frac{\partial}{\partial f_1}[f_1(1-q)\,\phi]-\frac{\partial}{\partial f_2}[f_2 q\phi]+\frac{\partial}{\partial f_3}[f_3(1-q)\,\phi]\right\} \\ & +Nc\left\{\frac{\partial}{\partial f_1}(D\phi)-\frac{\partial}{\partial f_2}(D\phi)-\frac{\partial}{\partial f_3}(D\phi)\right\} \end{aligned} \tag{6}$$

where, formally, in (6), p must be replaced by f_1+f_2, q by f_1+f_3 and D by $f_1(1-f_1-f_2-f_3)-f_2 f_3$. Thus, on a time scale proportional to N, the selection process is described completely by the initial conditions p_0, q_0 and D_0 and the parameters Nr, Ns, and Nc, and the chance of fixation at either locus is then a function of these alone.

A general solution of (6) has not been obtained though some results for Nr, $Ns < 0{\cdot}5$ can be given specifically in algebraic terms and we shall present later some results using matrix methods for $u(p_0)$ when $Nr < 0{\cdot}5$ but with no restriction on Ns.

Consider the rate of breakdown of linkage disequilibrium in small populations in the absence of selection. The recurrence equation for the mean value of D is then

$$D_t = (1-c)\,(1-1/2N)\,D_{t-1}$$

If c and $1/2N$ are small, so that their product can be ignored, we have

$$\begin{aligned} D_t &= (1-c-1/2N)\,D_{t-1} \\ &= D_0\,e^{-(2Nc+1)\,t/2N} \text{ approximately.} \end{aligned}$$

The half-life of the decline of the linkage disequilibrium coefficient to zero is given approximately by $t=1{\cdot}4N/(2Nc+1)$ generations. If Nr and Ns are small ($<0{\cdot}5$) it can be assumed that changes in the variance of gene frequency and in the disequilibrium coefficient will occur mainly as a result of genetic sampling and crossing-over and not as a result of selection. In any generation the expected change in p in any line is given by

$$\delta p = rp(1-p)/2+sD/2$$

and in q by

$$\delta q = sq(1-q)/2+rD/2.$$

We may assume, following Robertson (1960), that the average value of $p(1-p)$ will decline by a proportion $1/2N$ each generation and that the average value of D will similarly decline by a proportion $(c+1/2N)$. We have then for the expected total change in gene frequency

$$u(p_0) = p_0+Nrp_0(1-p_0)+NsD_0/(2Nc+1)$$

The expected change of gene frequency is then a linear expression in $2Nc/(2Nc+1)$. A linear relationship of change in gene frequency with this expression is in fact found in computer runs over a much wider range of Nr and Ns than that used in this derivation and this has very considerably simplified our discussion of the effect of linkage. If linkage is not initially at equilibrium, then the expected change in gene frequency may be greater or less than $2N$ times the change in the first generation, depending on the sign of the disequilibrium determinant.

Under the conditions of this derivation, segregation at a second locus has no effect on the chance of fixation of the first if linkage is in equilibrium at the start. We shall see later that, when we move to higher values of Nr and Ns, this is no longer true.

In most selection experiments, selection is for a quantitative character and changes in gene frequency are not directly observable. The selective advantages are then consequences of the effects of the loci on the character under selection. If these are small, we have approximately $r=i\alpha$, $s=i\beta$, where i is the selection intensity in standard units and α, β are the effects of the two loci on the metric character, expressed as the difference between the two homozygous genotypes divided by the phenotypic standard deviation, σ. Latter (1965*a*) has investigated the errors involved in this approximation. If considered in terms of the effect on changes in gene frequency, the errors appear to be compensatory in that the expression used above underestimates the selective advantage of genotypes with both positive and negative deviations from the population mean. If $i\alpha$ and $i\beta$ are small, additive action on the character under selection implies additive action on selective advantage, though this breaks down to some extent under intense selection, as we shall see later. The probable chance of fixation at the two loci can then be described in

terms of $Ni\alpha$, $Ni\beta$ and Nc and the consequent total change, R, in the population mean will be given by

$$R = \{\alpha[u(p_0)-p_0]+\beta[u(q_0)-q_0]\}\sigma \tag{7}$$

At any instant, the additive genetic variance can be expressed as

$$V_A = \sigma^2\{\tfrac{1}{2}\alpha^2 p(1-p)+\tfrac{1}{2}\beta^2 q(1-q)+\alpha\beta D\}.$$

This expression can be generalized to any number of loci with additive gene action and is then interesting in showing that, in the prediction of immediate response to artificial selection, the linkage disequilibrium need only be specified in terms of the disequilibrium determinants between the loci taken in pairs.

3. THE MONTE CARLO SIMULATION PROCEDURE

The simulation process was carried out on a high-speed computer, the I.C.T. Atlas. It was rather more abstract than that of other workers (Fraser, 1957; Martin & Cockerham, 1960; Gill, 1965; Latter, 1965*b*). Selection, recombination and sampling were all done at the gametic level and gametes were never paired into zygotes. Using the previous notation of gene effects, expressing all measurements in terms of the phenotypic standard deviation and taking the mean value of the genotype *aabb* as an arbitrary zero, the mean, m, of the population at any time is given by $m = p\alpha+q\beta$. Changes in gametic frequency are given by (4) and (5) with the selective values r and s replaced by $i\alpha$ and $i\beta$, and these equations include both the effect of selection and recombination. From the gamete frequencies so produced, the $2N$ gametes in the next generation were obtained by sampling from a multinomial distribution with parameters f_j by generating $2N$ uniform pseudo-random numbers X, $0 < X < 1$, and comparing each with the gametic frequencies. If $0 < X \leqslant f_1$, then a gamete AB was generated; if $f_1 < X \leqslant f_1+f_2$, then a gamete Ab was generated, and so on. Each of the parameters, N, $i\alpha$, $i\beta$, c, and the initial frequencies could be altered. In all runs, linkage equilibrium in the initial population was assumed. At the start of any run, the first step was one of selection by applying the above formulae to the initial frequencies, followed by the drawing of a random sample of gametes.

Each replication was continued to fixation or for $6{\cdot}25N$ generations, whichever occurred first. After this time, at least 99·9% of the total response at a single locus can be expected to be made if $Ni\alpha \geqslant 4$, 98·5% if $Ni\alpha = 2$, or 96·6% if $Ni\alpha = 1$. The average gene frequency at this time was then taken as the limit even if all lines had not reached fixation. Usually 400 replicates were run for each set of parameters. At fixation, the proportion of lines in which any allele is fixed is binomially distributed so that the standard error of the chance of fixation may easily be calculated. The chance of fixation at one locus when there was no segregation at the other was obtained by matrix iteration (Allan & Robertson, 1964), using the same population size as in the Monte Carlo runs rather than by using (2). This avoids small differences in the chance of fixation observed at a single locus when different population sizes are used for the same $Ni\alpha$ value (Ewens, 1963). These results for a single locus only

must also apply when the second locus has no effect on the character under selection or when Nc is very large, as in independent segregation of the two loci in a large population. In a very small population, for example $N = 8$, when the maximum biological value of Nc is 4, we have in fact detected some influence of independent segregation at the second locus on the chance of fixation of the first.

4. RESULTS

The outcome of any particular run is affected by five independently varying parameters, $Ni\alpha$ and p_0 referring to the first locus, $Ni\beta$ and q_0 to the second, and Nc. The output of any set of runs can be expressed in terms of the average chance of fixation at the two loci, $u(p_0)$ and $u(q_0)$, and the 'between line' disequilibrium determinant, calculated from the observed frequencies of fixation of the four gametes. It soon became clear to us that the results could be discussed most meaningfully in terms of the influence of segregation at a second locus on the chance of fixation at the first. The view of the results that we shall present here represents the combination of the Monte Carlo results with the insights we could gain into them by the application of algebra to the simpler situations.

We found no situations in which the chance of fixation at the first locus was significantly increased by simultaneous segregation at the second. We found none in which the between-line disequilibrium determinant was significantly positive at fixation and very many in which it was significantly negative.

(i) *The influence of the effect and initial frequency at the second locus*

Figures 1–4 have been chosen to illustrate various general aspects of the results. First we shall discuss the influence of changes in the parameters at the second locus. Concentrating on those situations in which there is no crossing-over ($Nc = 0$), segregation at the second has no detectable influence on the chance of fixation at the first until its effect is greater than one-half that of the first and, even when the gene effect is three-quarters that of the first, the influence on the chance of fixation is very small. We have found these conclusions to apply quite generally. An example is shown in Fig. 2. As the effect at the second increases further, the chance of fixation at the first passes through a minimum and then increases again. Figure 1 shows that the reduction is very dependent on the initial frequency of the preferred allele at the second locus. Clearly there has to be a minimum in this curve, as the second locus will have no influence when its initial gene frequency is zero or unity. The initial frequency at which segregation at the second produces the greatest reduction is dependent on the magnitude of its gene effect. We have found empirically that the minima in the chance of fixation $u(p_0)$, when plotted either against the gene effect or the gene frequency at the second locus, occur roughly when $Ni\beta q_0 = 0{\cdot}8$, whatever the parameters at the first. The chance of fixation of the preferred allele at the second is then also approximately 0·8. At this minimum, the reduction in the chance of fixation at the first increases as the gene effect at the second increases.

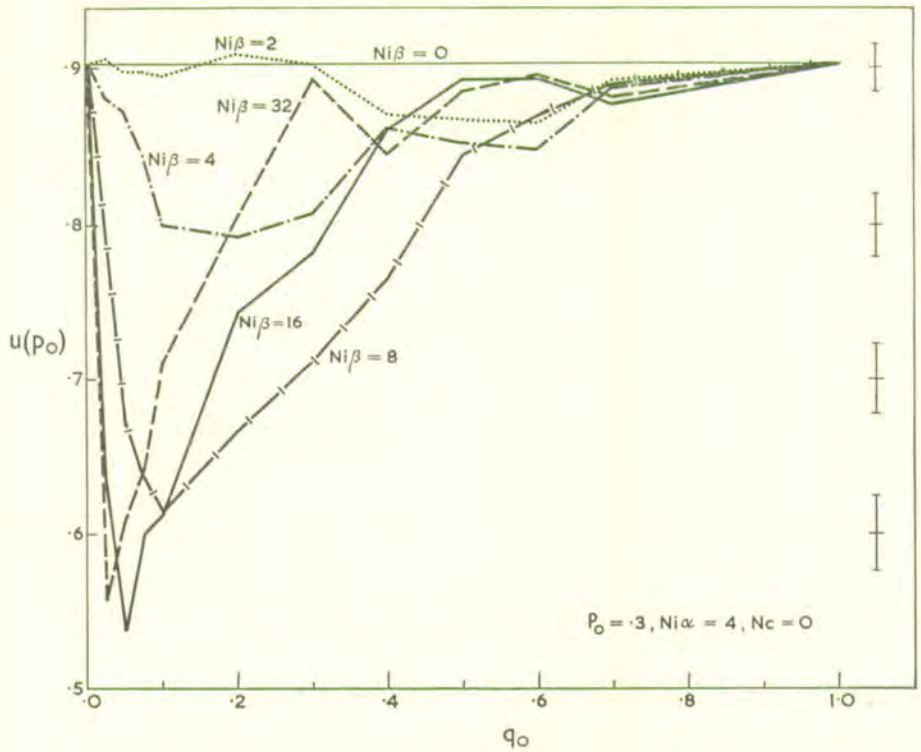

Fig. 1. The relationship between the chance of fixation at the first locus and the effect and initial frequency of the second. No crossing over. Typical ranges, of length two standard deviations, are shown.

(ii) *The influence of recombination frequency*

When $Ni\beta$ is small, the chance of fixation at the first locus is approximately linear in $2Nc/(2Nc+1)$ and this is well illustrated in Figs. 2–4. This expression goes from 0 to 1 as c increases from zero to infinity and the values $Nc=\frac{1}{4}$ and 1 divide this range into three equal intervals. Figure 4 shows that the curves for the three different crossover values are in fact equally spaced for all values of q_0, but, in Figs. 2 and 3, it will be seen that, although this prediction is reasonably satisfactory when $Ni\beta$ is less than 12, it obviously breaks down at higher values when the effect of increasing Nc from 0 is less than expected. In consequence, the value of $Ni\beta$ at which the minimum occurs is not independent of Nc and increases as the latter increases. At the high values of $Ni\beta$ the three curves become almost indistinguishable.

Runs not shown in these diagrams were made with a wide range of parameter sets ($p_0, q_0 = 0{\cdot}05$, 0·1, 0·3, 0·5 and 0·7; $Ni\alpha = 2, 4, 8$ and 16, and either $Nc = 1, \frac{1}{4}$ and 0, or $Nc = 4, 1, \frac{1}{4}, \frac{1}{16}$ and 0). For each set of the other four parameters, the linear regression of $u(p_0)$ against $2Nc/(2Nc+1)$ was calculated, the line being forced through the matrix iteration result for $Nc=\infty$. It was found that 97·4% of the variation in $u(p_0)$ between different Nc values could be removed by the linear regressions.

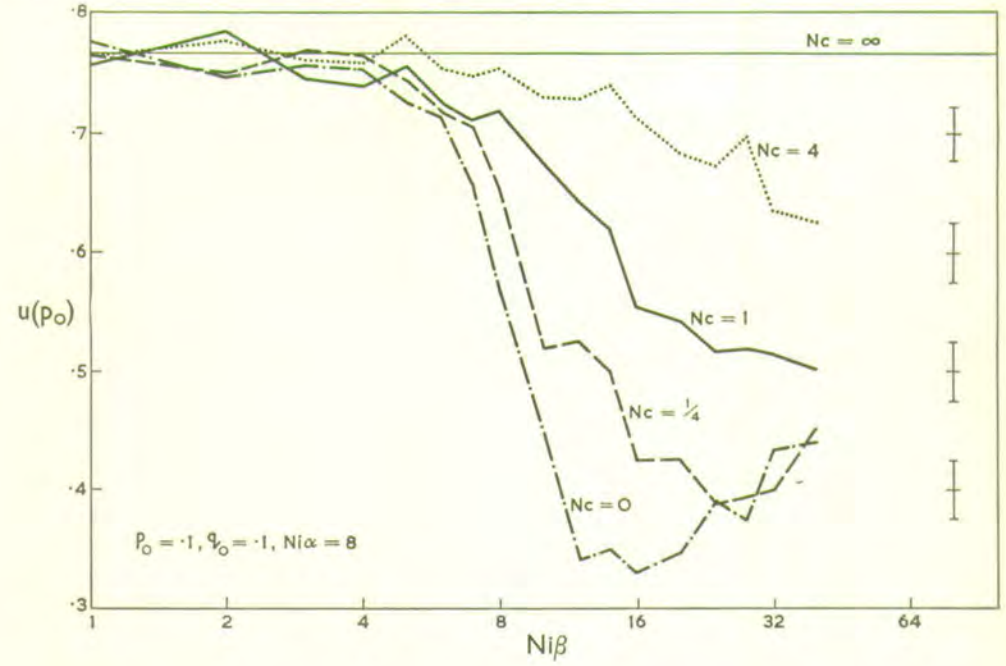

Fig. 2. The relationship between the chance of fixation at the first locus and the effect at the second, for various recombination values. Typical ranges, of length four standard deviations if $Ni\beta \leqslant 8$, or two standard deviations if $Ni\beta > 8$ are shown.

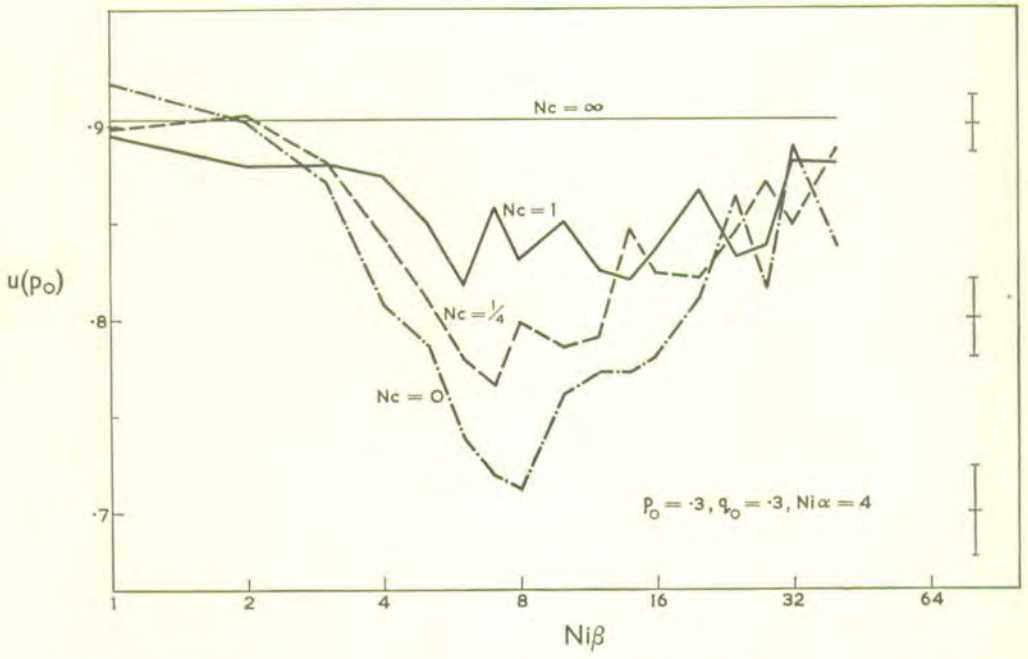

Fig. 3. As Fig. 2, but with the effect halved at the first locus. Typical ranges of length two standard deviations are shown.

Nevertheless the residual variation due to curvilinearity was highly significant in many cases.

Figure 4 also shows the effect of altering population size in the computer runs for fixed values of $Ni\alpha$, $Ni\beta$ and Nc. The curves for a population size of 8 are indistinguishable from those with a population size of 16.

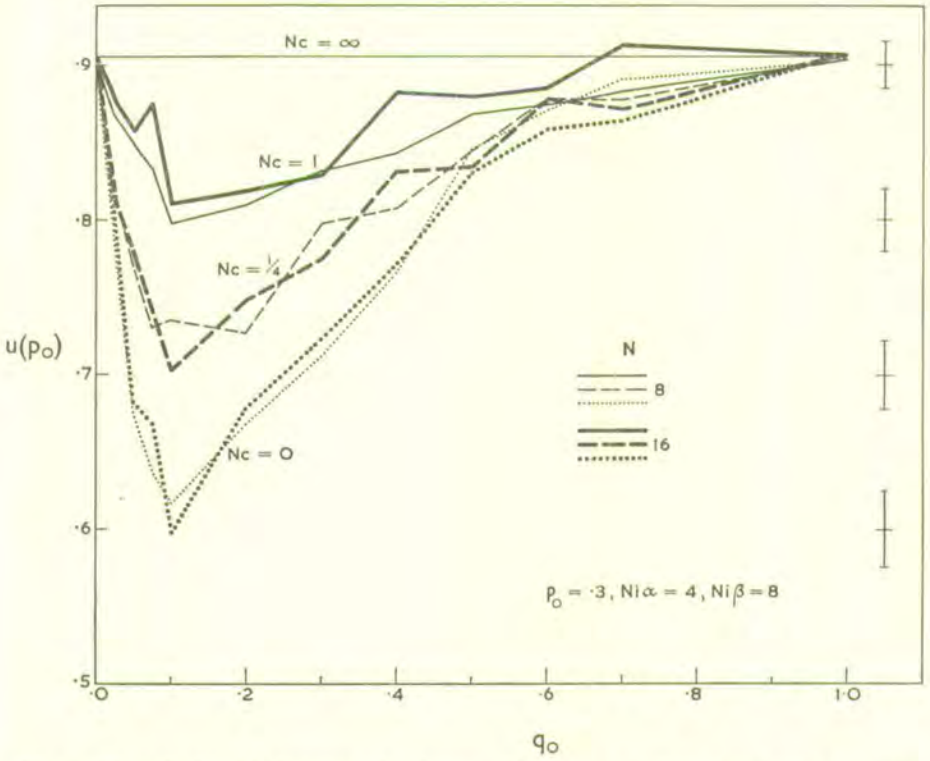

Fig. 4. The relationship between the chance of fixation at the first locus gene and the initial frequency at the second for various recombination values. Estimates were made at two levels of population size. Typical ranges, of length two standard deviations, are shown.

(iii) *Changes in the parameters at the first locus*

Any discussion of the influence of changes in the parameters at the first locus is complicated by the fact that in the absence of segregation at the second, variations in these will affect the chance of fixation. We are then concerned to find a description of the effects of this segregation on the chance of fixation which will be as far as possible independent of the parameters at the first locus. Segregation at the second reduces the chance of fixation at the first. This can be thought of as a reduction of the effective selection intensity at the first locus. From each computer run, we calculated from Kimura's formula (2) the effective value of $Ni\alpha$ (denoted $\widehat{Ni\alpha}$) which, from the given initial gene frequency, would give the observed chance of fixation if the first alone was segregating. Figure 5 gives examples of the use of this

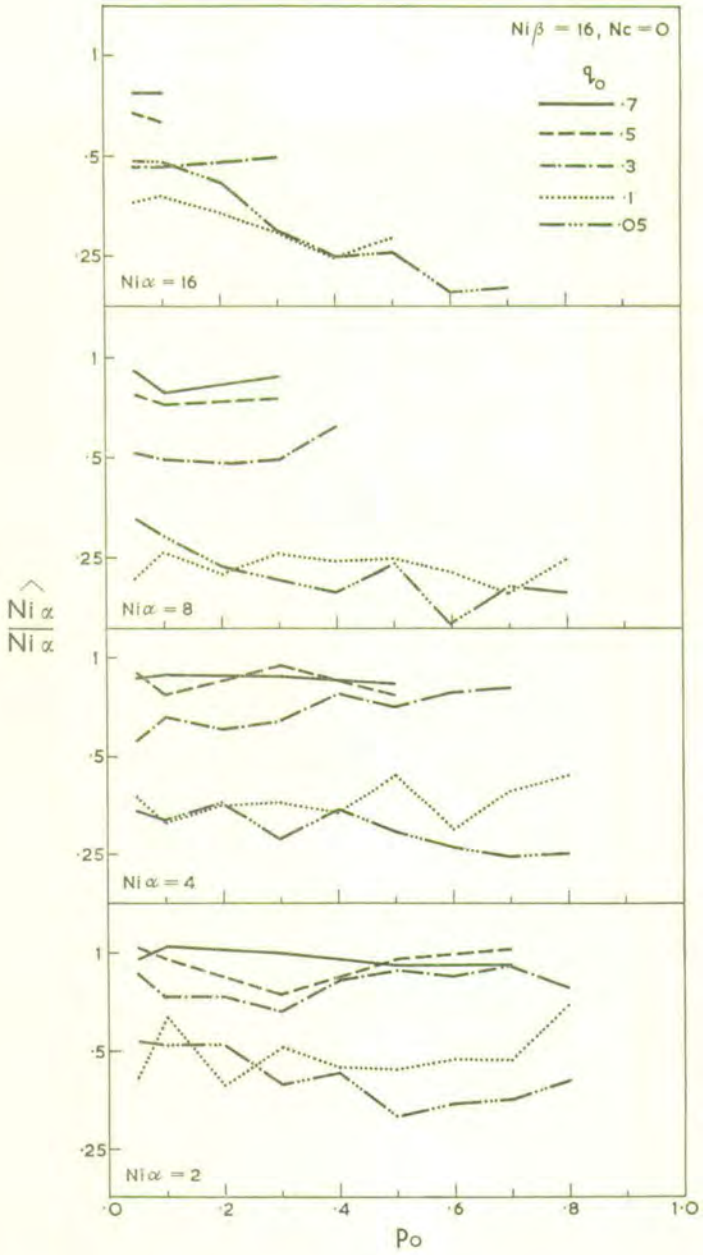

Fig. 5. The effective selection parameter, $\widehat{Ni\alpha}/Ni\alpha$, at the first locus as influenced by segregation at the second ($Nc = 0$).

transformation in evaluating the interaction of $Ni\alpha$ and p_0 with the other variables. Because the sampling variance becomes very high as $u(p_0)$ approaches unity, no points are plotted when the observed value exceeds 0·99. It is quite clear that the effect of the segregation at the second locus, if expressed in this way, is almost independent of the gene frequency at the first for values of $Ni\alpha$ up to about 4.

However, this independence breaks down at low values of q_0, when $\hat{Ni\alpha}$ is reduced as p_0 increases.

It is a necessary consequence of the theoretical model of the process which will be presented in a subsequent paper that $\hat{Ni\alpha}/Ni\alpha$ will be independent of both p_0 and $Ni\alpha$ when $Ni\alpha < 0{\cdot}5$. However, it will be seen that as $Ni\alpha$ increases, the observed value of $\hat{Ni\alpha}/Ni\alpha$ declines. An example is given in Table 1. As $Ni\alpha$ increases still further to values greater than $2Ni\beta$, when $u(p_0)$ will cease to be affected by the segregation at the second locus, $\hat{Ni\alpha}/Ni\alpha$ must obviously approach unity.

Table 1. *The relationship between* $\hat{\text{Ni}\alpha}/\text{Ni}\alpha$ *and* $\text{Ni}\alpha$ *for a model with* $\text{Ni}\beta = 16$, $\text{q}_0 = 0{\cdot}1$, $\text{Nc} = 0$, *averaged over a range of* p_0 *from* 0·05 *to* 0·8

$Ni\alpha$	$\hat{Ni\alpha}/Ni\alpha$
0	0·56
2	0·49
4	0·37
8	0·24
16	0·32

(iv) *The rate of selection advance*

We have so far only discussed the final chance of fixation at the two loci. Typical response curves are shown in Fig. 6, which give the smoothed averages of 3200 Monte Carlo replications with $N = 8$ for $Nc = 1$, $\frac{1}{4}$ and 0 respectively. The results for $Nc = \infty$ were obtained by iteration of the matrix of transition probabilities for a single locus. Clearly in the first few (say, $N/2$) generations, linkage has little influence on the rate of response, but then with tight linkage the latter rapidly slows down. After about $2N$ generations, the response has almost ceased for both $Nc = 0$ and $Nc = \infty$ but there is continued response for the two intermediate frequencies of crossing over. Since the approach to the limit is asymptotic, Robertson (1960) used the half-life of the selection process, the time taken for the mean gene frequency to get half-way to the limit, as a measure of the time scale of the response. Approximate half-lives for the example of Fig. 6 are shown in Table 2. It can be seen that as it is only the response in later generations which is reduced by tight linkage, the half-life is reduced at the lower values of Nc.

Latter (1966*a*) gives further results for the case of equal initial frequencies and selective advantages with two loci. He finds that while the half-life of the selection process is reduced the time taken to obtain 95% of the total advance is usually increased with intermediate recombination values, because of the prolonged period of late response.

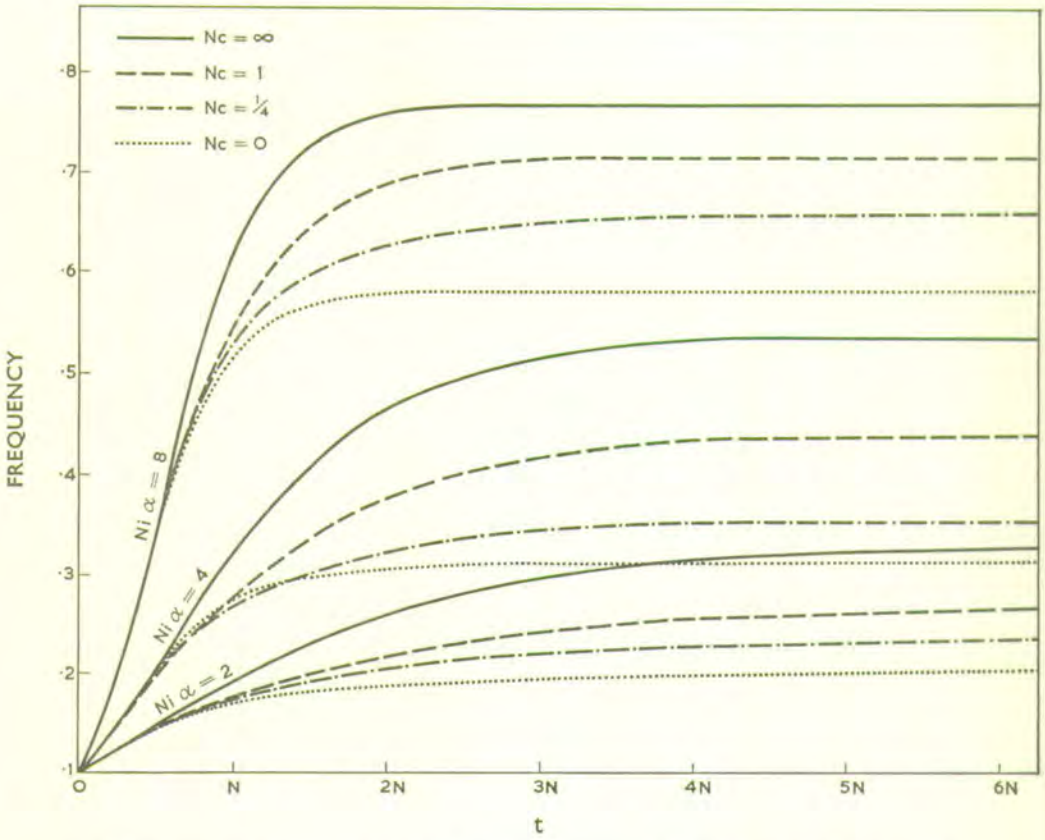

Fig. 6. Response curves at the first locus as influenced by its effect and tightness of linkage to the second. Time is measured in generations.

Table 2. *Half-lives* ($\times$ N *generations*) *of the selection process for* $p_0 = q_0 = 0{\cdot}1$ *and* $Ni\beta = 8$

	Nc			
Niα	∞	1	¼	0
2	1·31	1·19	0·86	0·65
4	1·00	0·95	0·66	0·57
8	0·64	0·62	0·57	0·50

Another view of the effect of tight linkage is given in Fig. 7, in which the mean value of p after varying numbers of generations is plotted against q_0. The effect of the segregation at the second locus is seen as the depression of p at low values of q_0. Before $N/2$ generations, this segregation has no effect on the gene frequency at the first, but at N generations p, at the value of q_0 which has maximum effect, is below that at other values and there is little change in p after this point. The diagram shows why the half-life of the process is reduced when the preferred allele at the second locus is at its most effective frequency. Examination of the curves for

$Ni\beta = 8$ and 16 in Fig. 7 shows that at $q_0 = 0{\cdot}2$, both the total response and the half-life are greater for the higher value of $Ni\beta$.

Figure 8, which is of the same kind as Fig. 7, shows the effect of variation in $Ni\beta$ on the mean values of p at different times and includes curves for three values of Nc. When $Ni\beta$ has its maximum effect on $u(p_0)$ at the given initial frequency ($Ni\beta = 8$) there is again little effect of linkage in less than $N/2$ generations, but as $Ni\beta$ increased further it will be seen that almost all the reduction in response due to tight linkage now occurs in the earlier generations. This is to be expected since, for such high

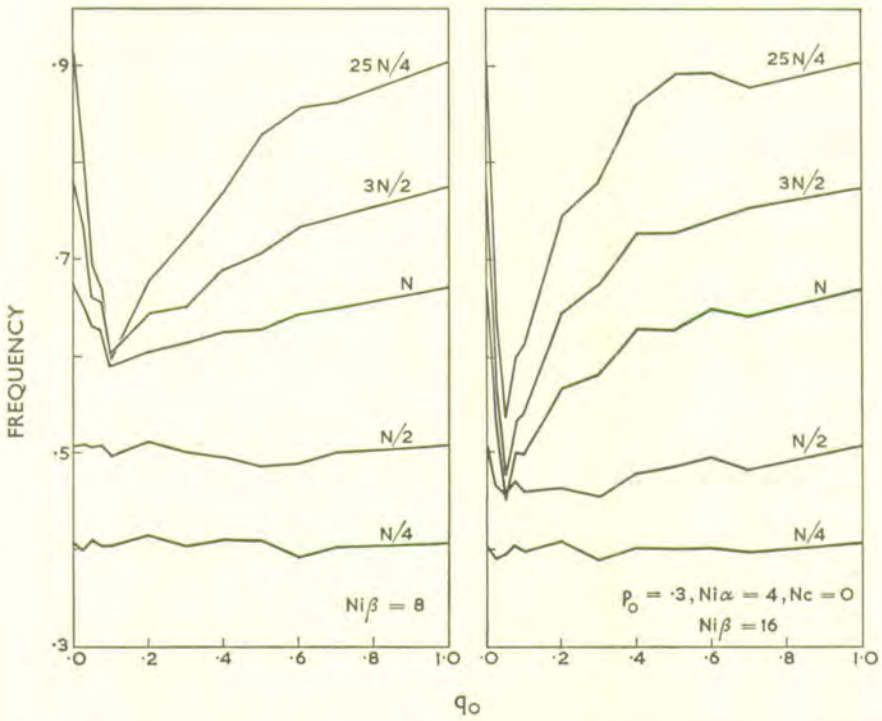

Fig. 7. The average frequency at the first locus at various times during the selection process, measured in generations, as influenced by the initial frequency and effect at the second.

values of $Ni\beta$, the second locus becomes fixed very quickly and only during this period is there segregation at both loci. It can be shown, by iteration of the transition probability matrix, that for a single gene with $Ni\beta = 32$ and $q_0 = 0{\cdot}3$, 99% of the expected change in gene frequency has been made in the first $0{\cdot}33N$ generations, whereas for a gene with a much smaller effect ($Ni\beta < 0{\cdot}5$) it takes $4{\cdot}61N$ generations for this point to be reached. When one locus goes to fixation so quickly it is clear that crossing-over has very little time to affect the outcome. In Fig. 8 it can be seen that at high values of $Ni\beta$, no more progress is made with $Nc = 1$ than with $Nc = 0$. With smaller values of $Ni\beta$, however, there is more time for recombination to occur. The shortened period of response when $Ni\beta$ is high then provides an explanation

of the unexpectedly small effect of the relaxation of linkage at high $Ni\beta$ values in Figs. 2 and 3.

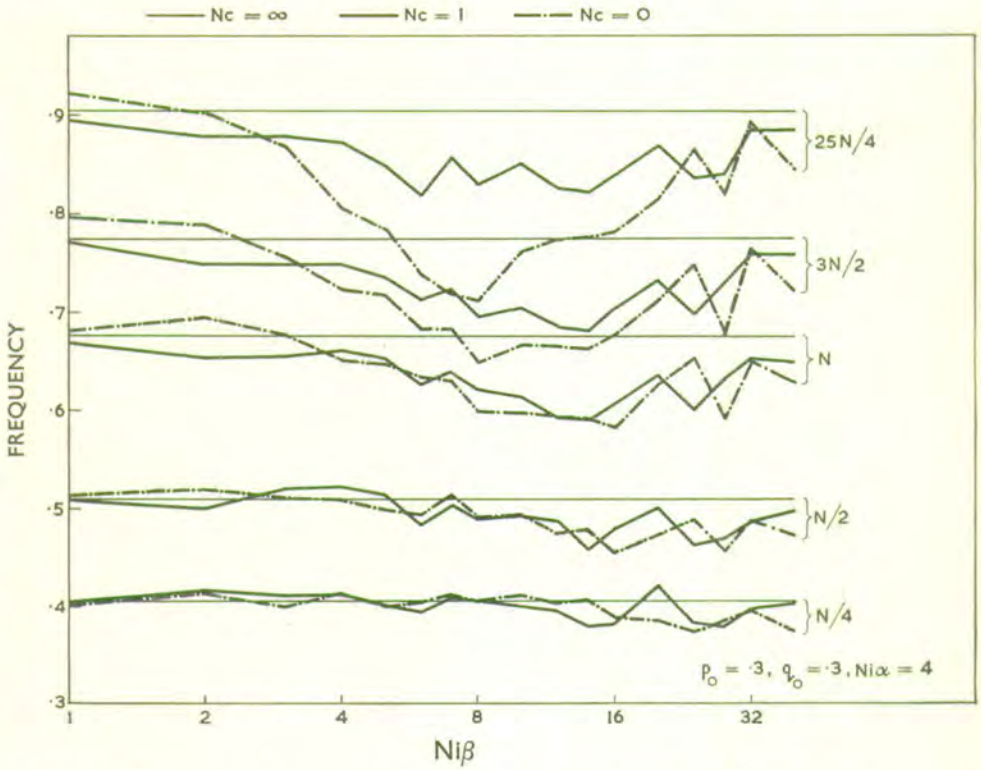

Fig. 8. The average frequency at the first locus at various times during selection, as influenced by the effect and tightness of linkage with the second.

(v) *The chance of fixation of the different gametes*

We have so far considered only the chance of fixation of the individual alleles; we shall now discuss their joint chance of fixation. Figure 9 shows, for $p_0 = q_0 = 0{\cdot}1$, $Ni\alpha = 8$, the effect of variation in Nc and in $Ni\beta$. Of these diagrams two are chosen so that $\beta \leqslant a/2$, one so that β is almost as large as α, and the final diagram shows the case of equal effects. Data from these runs have also been seen in Fig. 2. In Fig. 9 the results are plotted against $2Nc/(2Nc+1)$. As would be expected, at the lower values of β only the chance of fixation at the locus with the smaller effect is reduced as linkage becomes tighter. When the two effects are equal, the chance of fixation of the preferred alleles is reduced at both loci by tight linkage.

Latter (1965*b*) has shown that with equal effects at the two loci the chance of fixation of the unfavourable coupling gamete, *ab*, is not influenced by the degree of recombination and we find in Fig. 9 that this result holds even when the effects are unequal. The chance of fixation of the gamete aB is affected by linkage only as β approaches α. On the other hand, the chances of fixation of the gametes AB and Ab are influenced by the tightness of the linkage in all the cases. When $\beta \leqslant \alpha/2$, the favourable coupling gamete AB is less frequently fixed and the repulsion gamete

Ab more frequently fixed with tight linkage, in such a way that the sum of their frequencies is not affected. So, if one gene has a much smaller effect than the other, the reduction in its chance of fixation as linkage becomes tighter takes place only amongst gametes in which the preferred allele at the other locus is fixed. This is to be expected in view of the results in the previous section. The gametes *ab* and *aB* are most likely to be fixed in the early generations of the selection process before the tightness of the linkage much affects it.

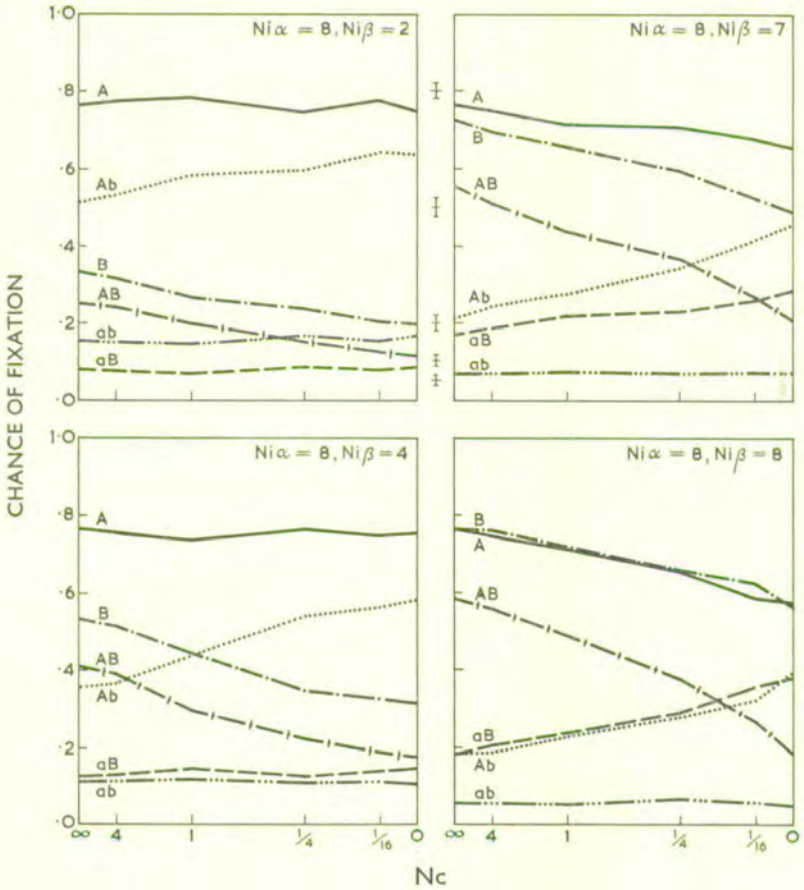

Fig. 9. The chance of fixation of the favourable alleles and the four gametic types with initial frequencies $p_0 = q_0 = 0{\cdot}1$.

We see in Fig. 9 that, as linkage becomes tighter, the chances of fixation of the repulsion gametes *Ab* and *aB* either remain constant or increase, that of *ab* remains constant and that of *AB* is reduced. As a consequence there is a negative disequilibrium D_L between lines at the limit, where

$$D_L = u(AB)\,u(ab) - u(Ab)\,u(aB)$$

where $u(-)$ is the chance of fixation of the specified gamete. In Fig. 9 with $Nc = 0$, the values of D_L are $-0{\cdot}0351$, $-0{\cdot}0667$, $-0{\cdot}1129$ and $-0{\cdot}1383$ when $Ni\beta = 2$, 4, 7 and 8 respectively. In general in our computer runs we have found an excess of

repulsion gametes at the limit. Of the 210 runs having 400 replicates with the wide range of parameter sets mentioned earlier, D_L was zero in 72 cases (because a particular allele was fixed in all replicates), it was negative in 130 and positive in only 8. In none of the latter did D_L differ significantly from zero at the 5% level. Similarly, the observation that the chance of fixation of the *ab* gamete was not altered by the degree of linkage was found to hold at all levels of effects. Where the gene effects differed by a factor of at least 2, it was generally found that the chance of fixation of the repulsion gamete containing the unfavourable allele at the locus with the larger effect was little affected by the tightness of linkage.

(vi) *Change in the population mean under artificial selection*

We have discussed the results so far in terms of the chance of fixation of the individual gametes, but in a selection experiment for a quantitative character all that can usually be observed is the change in the population mean. This is a function

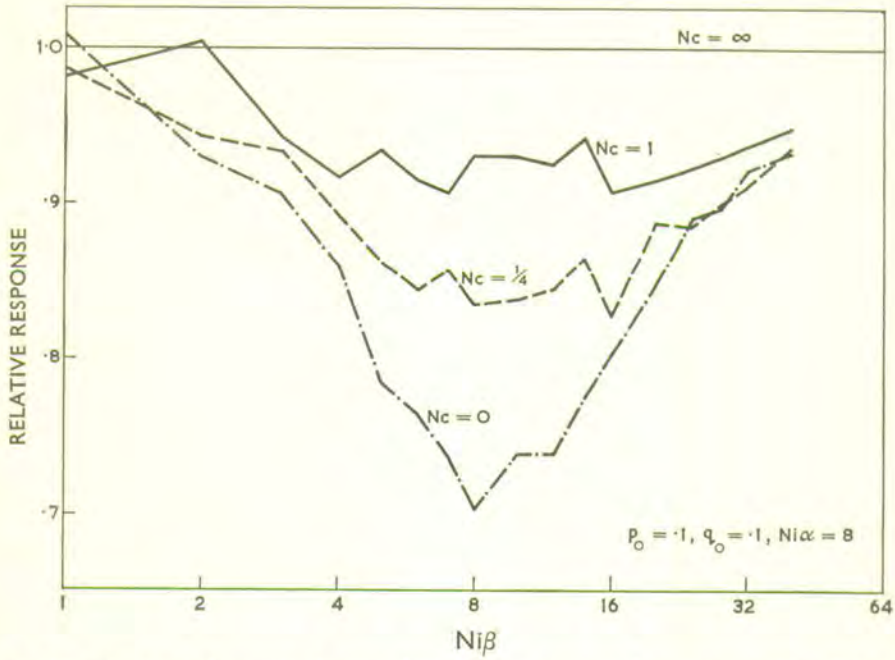

Fig. 10. The total change in the population mean expressed as a proportion of that expected from independent genes with the same effects and initial frequencies.

of the effects and changes in frequency at all loci contributing to the trait and in our case will be given by (7). To compare results from different initial frequencies and effects, we shall consider the response R observed for some parameter set as a proportion of that expected from the same set with free recombination between the loci. The greatest proportional reductions in R caused by tight linkage are found when α and β are approximately equal. An example is shown in Fig. 10, in which $Ni\alpha$ is kept constant and $Ni\beta$ is varied. The minimum of the curve of relative response occurs when the effects are approximately equal at the two loci. This

result could have been anticipated from the earlier data for, in a model in which one locus has a much larger effect than the other, it has been shown that the change in gene frequency at the former (which will contribute most to changes in the mean of the population) is scarcely influenced by the smaller linked gene. Thus, the response in the mean will not be much influenced by the tightness of linkage when the genes have widely unequal effects on the quantitative trait.

The greatest reduction in response with tight linkage occurs when both genes have a low initial frequency and large effect. There are two reasons for this. We have to consider the same locus both as influencing the other one and being influenced by it. We showed earlier that the effect of one locus on another can best be expressed in terms of the proportional reduction in the effective value of $Ni\alpha$, and that this occurs when $Ni\beta q_0$ is in the region of 0·8, when the chance of fixation of the B allele is itself about 0·8. For the effects to be perceptible, $Ni\beta$ should be greater than 2. Now consider the sensitivity of the second locus to the segregation at the first. We can consider this as the proportional change in advance under selection, $u(q_0)-q_0$, for a given proportional change in $Ni\beta$. If the values of $Ni\beta$ are sufficiently large that we can ignore the denominator in equation (2), it can be shown that the sensitivity is at a maximum when $Ni\beta q_0(1-q_0)=\frac{1}{2}$. When $q_0=0{\cdot}2$, the maximum sensitivity will be achieved when the chance of final fixation (given by $Ni\beta=3{\cdot}125$) is 0·71. As q_0 declines, the chance of fixation for maximum sensitivity approaches the value of 0·64. Thus, for maximum influence we require a value of $u(q_0)$ of 0·8 and for maximum sensitivity we require a value slightly more than 0·64. It is not surprising then that Latter (1965*b*) found, when investigating two loci with equal effects and equal gene frequencies, that tight linkage had most effect on the advance under selection when $(u(q_0)-q_0)/(1-q_0)$ was in the region of 0·7.

It is sometimes possible in artificial selection programmes to vary the effective amount of crossing-over. One could, for instance, insert between each generation of selection a generation of relaxation with a large number of parents. This would effectively double the value of c in our equations. It is therefore of interest to know what effect this would have on the selection advance. In the situation in which linkage has its greatest effect (see Fig. 8) there appears to be an almost linear regression of change in gene frequency on $2Nc/(2Nc+1)$ in that the values for $Nc=\frac{1}{4}$ and 1 are equally spaced between $Nc=0$ and ∞. With $Nc=0$, about 30% of the total advance is lost. Assuming the linear relationship on $2Nc/(2Nc+1)$ to hold exactly, the expected responses with $N=20$, expressed as a proportion of the advance with free recombination, would be as follows:

Cross-over frequency c	*Proportional advance*
1/160	0·771
1/80	0·811
1/40	0·862
1/20	0·913
1/10	0·953
1/2	1·000

Doubling the recombination fraction produces at most an increase of 6% in the advance under selection. This occurs at $2Nc = 1$, when the curve of $2Nc/(2Nc+1)$ against $\log Nc$ has its greatest slope.

These results have some bearing on the intensity of artificial selection which should be applied in order to maximize the advance. In a mass selection programme, the number of individuals that can be measured in any generation (T) can be regarded as fixed. If selection affects only a single locus, or several independently segregating loci, it can be shown that the advance will be a function of Ni, where N is the number of animals selected to be used as parents and i is the selection intensity in standard units. This is at a maximum when the proportion of individuals selected is 0·5 (Dempster, 1955; Robertson, 1960), and the advance is symmetrical for variation about this value in the proportion selected. When two linked loci are under selection it might be expected that for two values of the proportion selected (say, 0·4 and 0·6) which give the same value of Ni, the selection advance would be greater for that with the lesser intensity of selection because Nc will then be greater. The advance under selection will no longer be symmetrical about $N/T = 0{\cdot}5$. In Table 3 we have therefore chosen for consideration a situation in which this effect should be most easily detected, i.e. two loci with equal effects on the character under selection at initial frequencies chosen so that the effect of linkage will be at its maximum and the linkage distance chosen so that the advance will be most sensitive to changes in Nc ($T = 40$, $\alpha = \beta = 0{\cdot}5$, $p_0 = q_0 = 0{\cdot}1$, and $c = 0{\cdot}025$). Figure 10 shows that in this situation the advance is almost linear on $2Nc/(2Nc+1)$. We have therefore used this relationship for interpolation of our Monte Carlo data.

Table 3. *Chance of fixation of an additive gene when* 40 *individuals* (T) *are recorded,* $\alpha = \beta = 0{\cdot}5$ *and* $p_0 = q_0 = 0{\cdot}1$

	Proportion selected								
	0·05	0·1	0·25	0·4	0·5	0·6	0·75	0·9	0·95
No linkage	0·34	0·51	0·71	0·78	0·80	0·78	0·71	0·51	0·34
$c = 0{\cdot}025$	0·31	0·46	0·61	0·66	0·70	0·70	0·65	0·49	0·33
$c = 0$	0·30	0·45	0·52	0·60	0·61	0·60	0·52	0·45	0·30

Both when the genes are segregating independently and when there is no recombination, the expected selection advance will be proportional to Ni and will be symmetrical about $N/T = 0{\cdot}5$. The second line of the table shows that, when $c = 0{\cdot}025$, the maximum in the chance of fixation, considered as a function of N/T, is only slightly shifted and occurs when N/T is about 0·55. Considerations of linkage should not greatly influence the intensity of selection to be practised if only two loci are involved. However, more drastic effects might be found with more than two.

5. DISCUSSION

There may appear to have been some contradiction between our earlier theoretical discussion and the Monte Carlo results. We stated that, when both $Ni\alpha$ and $Ni\beta$ were small, the expected advance under selection could be specified in terms of the

initial gene frequencies and the initial disequilibrium determinant and the distance between the two loci appeared only in the term containing the latter. Nevertheless, in the Monte Carlo studies it appeared that, even though we start with linkage initially at equilibrium, the advance under selection is dependent on the tightness of linkage between the two loci. How has this come about?

Felsenstein (1965) has presented a discussion of the effect of selection on linkage in infinite populations. He points out that if the genes concerned affect fitness in a multiplicative manner (i.e. if w_1, w_2, w_3 and w_4, are the relative fitnesses of the AB, Ab, aB and ab gametes and $w_1 w_4 = w_2 w_3$) then an infinite population in initial linkage equilibrium will remain in equilibrium during selection. He points out that truncation selection on a metric character will generally lead to immediate linkage disequilibrium. Nei (1963) showed that a large population initially in linkage equilibrium exposed to truncation selection has in the first generation a disequilibrium determinant given in our terminology by

$$D_1 = -\tfrac{1}{4} i^2 \alpha\beta p(1-p)\, q(1-q)$$

This formula assumes that the genes are acting additively on the character under selection. It is in fact only an approximation and inclusion in the expressions for selective advantages of squared terms in the gene effects leads to the expression

$$D_1 = \tfrac{1}{4}(ix - i^2)\, \alpha\beta p(1-p)\, q(1-q)$$

where x is the truncation point in standard units. Since $ix - i^2$ is always negative, a negative disequilibrium will be set up and the rate of response will therefore be reduced by tight linkage. In our case we have assumed an additive combination of the genes at the different loci in their effect on the fitness of the four gametes. Such selection will certainly lead to some negative disequilibrium in a large population and we decided to investigate whether this was responsible for the effect of linkage on selection limits in our case. We therefore set up for some values of the parameter sets a system of multiplicative selective advantage of the gametes. Such a modification is not as simple as it sounds, as starting from a given initial gene frequency, we wish to have the same chance of fixation in both cases with free recombination between the loci. A comparison of the additive and multiplicative model was run for a total of 80 different parameter sets and the results showed little, if any, difference in chance of fixation with tight linkage in the two models. The differences between the additive and multiplicative models obviously depend on the range of variation in the selective values of the different gametes. Although we used values for $i\alpha$ and $i\beta$ as large as one (a magnitude which would very rarely be encountered in practice) no differences were obtained between the two models. We therefore conclude that the reductions in chance of fixation with our model are not due to any great extent to a build-up of negative disequilibrium due to selection alone, predicted by Felsenstein's equation.

The solution to this problem comes from an examination of the effect of multiplicative selection when the disequilibrium determinant is not zero. It can then easily be shown that

$$D_1 = k_1 D_0 / (1 + k_2 D_0)^2$$

where k_1 and k_2 are functions of the gene frequencies and selection coefficients and are always positive. If we now consider the joint effect of genetic sampling and multiplicative selection, we see that in the first finite samples taken from the population initially at equilibrium, D will be distributed about a mean of zero with a variance depending on the sample size. After multiplicative selection, in which the D distribution will be modified according to the above formula, the average value will now be negative. A consideration of our computer runs would suggest that, even with multiplicative action, the mean negative disequilibrium determinant decreases as the square of time in the early generations, passes through a minimum and then rises to zero at final fixation.

We have not found the analysis of this process in terms of the development of the disequilibrium determinant during selection particularly illuminating and have come rather to a view of the situation in terms of the effective population size in which gene frequency changes at the locus with the smaller effect take place. This view will be given a mathematical treatment in a subsequent paper, but we may well sketch it out here for tight linkage. Consider a situation in which the B allele is at low frequency in the initial population but in which the selection process is such that, if in the initial sample there is a gamete containing B, it will almost certainly be fixed. There will then be two kinds of initial samples. In the first, no gametes containing B will be present and the expected change in p will be that calculated from equation (2). The other kind of initial sample will contain very few gametes containing the B allele. These will spread rapidly through the population under selection. With tight linkage, the change in frequency of the A allele in such lines has to take place within a population of gametes which may be very small in the early generations though, of course, as B becomes fixed it will approach $2N$. We may then expect that the average change of gene frequency at the A locus will be less in those lines in which the B gene becomes fixed and that tight linkage will therefore reduce the overall chance of fixation of A. As the initial frequency of B decreases, we have two opposing effects which lead to the minimum in the curve in Fig. 1. The first consequence will be a reduction of the number of B alleles in the initial sample (thus reducing the chance of fixation of A) until this effect is overcome by an increase in the proportion of initial samples contain no B alleles at all (so increasing the chance of fixation of A).

From this way of visualizing the problem, we can also obtain insight into some of the other surprising results. We have said that when segregation at the second locus has its greatest influence, only the changes of gene frequency at the first locus among gametes containing the desirable allele at the second are of importance in determining the final chance of fixation. The number of such gametes may be very small in the early generations of the selection process. Consider the situation in which almost all initial samples contain at least one gamete with the desirable allele at the second locus. If we then double the effect at the second locus, such gametes will increase in frequency more rapidly and as a consequence, the effective population size within which frequency changes at the first locus have to take place will be small for a shorter period of time. The expected change in gene frequency at the

first locus is then increased by increasing the effect at the second locus. We have here then an explanation of the minima in the curve of $u(p_0)$ plotted against $Ni\beta$, which we showed in Figs. 2, 3 and 4.

Now consider the effect of an increase in the gene effect at the first locus. Table 2 then shows that a greater part of the advance under selection will take place in the early generations but it is precisely in these early generations that the effective population size, with respect to changes in gene frequency at the first locus, is at its smallest. It then follows that the relative effect on the chance of fixation at the first locus will be greater as its own effect increases. This will hold until the latter approaches the same size as the effect at the second locus. This then provides us with a satisfactory explanation of the minimum value of $\hat{Ni\alpha}/Ni\alpha$ found in Table 1.

Latter (1966*b*) has discussed in some detail the interaction of linkage intensity and population size, using computer simulation with two additive loci with equal gene effects and initial gene frequency. He concentrated attention on the situation in which he had found that the restrictive effect of linkage was greatest, i.e. when, under free recombination, $u(p_0)-p_0=0{\cdot}70(1-p_0)$. We were interested to see to what extent the interactions might be removed when linkage intensity was measured on the scale $2Nc/(2Nc+1)$. At his lower population sizes ($N=5$ and 10) the regression of response in the transformed linkage value was reasonably linear but this was clearly not so for the higher values ($N=20$ and 40). There was then a higher chance of fixation for intermediate values of c than would be expected from a linear relationship, i.e. the curve was concave downwards. This is opposite to the curvilinearity we found for high values of $Ni\beta$ when $\beta \gg \alpha$ (see Fig. 3).

Latter's experiments at the higher two population sizes correspond in our notation to runs with $Ni\alpha=Ni\beta=18$, $p_0=q_0=0{\cdot}035$ and $Ni\alpha=Ni\beta=36$, $p_0=q_0=0{\cdot}017$ respectively, somewhat higher values of Ns than we have dealt with. However, it is interesting that the curve of the chance of fixation of the AB gamete is concave downwards in our Fig. 9 when $Ni\alpha=Ni\beta=8$, $p_0=q_0=0{\cdot}1$, our most comparable experiment. In his theoretical treatment of the results, Latter lays particular stress on that phase of such selection runs in which only the equivalent gametes Ab and aB are segregating, a phase in which no selection is taking place even though the original $Ni\alpha=Ni\beta$ values were high. This phase is ended either by random fixation of one of the two or by the production of an AB gamete by crossing-over.

We would suggest that such a situation is a very special case due to the equality of gene effects at the two loci. With two alleles segregating and low selection pressures, it is known that the half-life of the process is $1{\cdot}4N$ generations. Selection reduces this by a factor which is dependent on Ns, where s is the difference in selective advantage between the two. Figure 11 (an extension of Fig. 3 in Robertson (1960)) shows the magnitude of this reduction. At higher values of Ns, the half-life at a given initial frequency is proportional to $1/Ns$. In the nomenclature of this paper, Ns is equal to $Ni(\alpha-\beta)$. If a value of $Ni(\alpha-\beta)$ of four can greatly reduce the period of joint segregation of the Ab and aB alleles, when $Ni\alpha$ and $Ni\beta$ are of the order of 40, we would suggest that we are here dealing with a very special case which

would be much altered at the higher population size by a relative difference of only 10% between the gene effects at the two loci. This should perhaps suggest caution in generalizing too much from selection simulation studies on models in which all loci have equal effects.

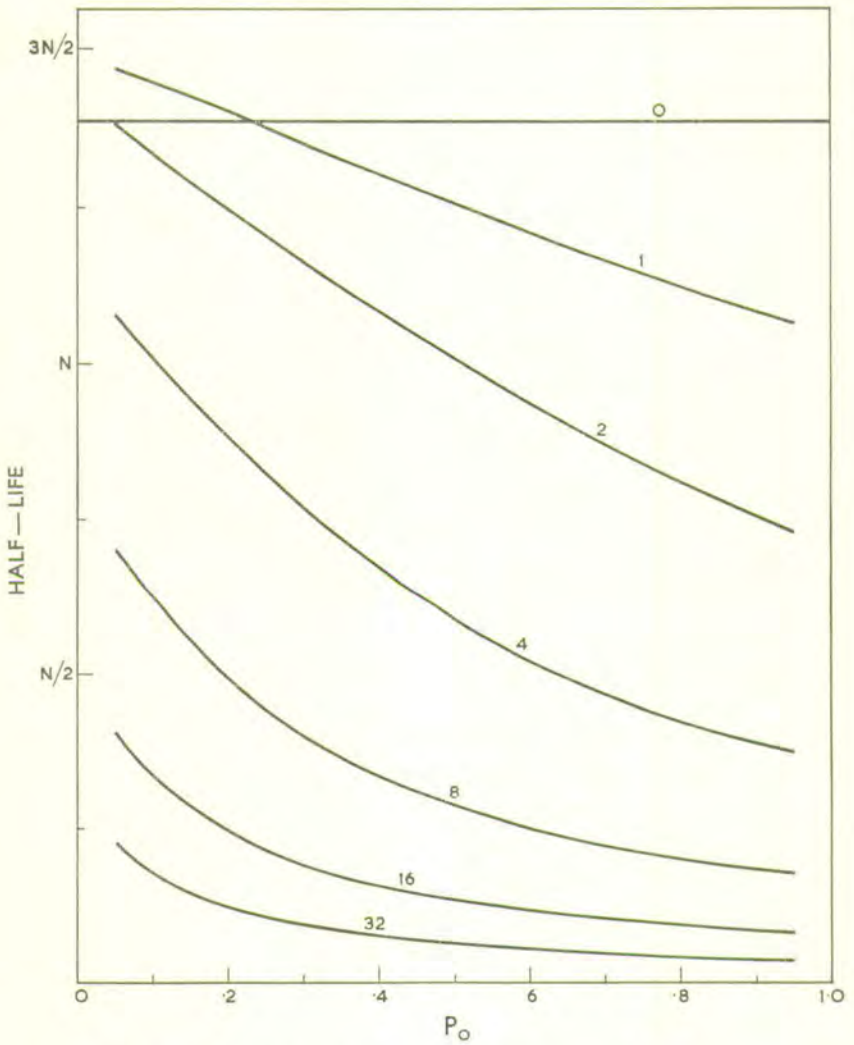

Fig. 11. The half-life of the selection process when two alleles are segregating at one locus.

We should now turn to some of the assumptions and limitations of this study. From the diffusion equation, it was argued that computer runs need only be made at one level of population size but the parameters $i\alpha$, $i\beta$ and c used were frequently much larger than those required for the diffusion approximation to hold. Nevertheless our results, including those of Fig. 4, indicate that the use of $Ni\alpha$, $Ni\beta$ and Nc as sufficient parameters is highly robust against departures from the underlying assumptions. Again, some approximations were made in the simulation procedure,

partly to reduce computing time. In particular, the algebra developed for infinite populations which was used to simulate selection and recombination entirely in terms of gametes, assumes that Hardy–Weinberg equilibrium holds and also that there is no distinction between the sexes and that self-fertilization is permitted. Errors introduced by these approximations will become smaller, the larger the population size used, but small N values were usually run to minimize computation. A similar kind of inaccuracy was introduced in the definition of the selective advantage in terms of the effects of the genes on the character under selection, which are precise only for genes of small effect. Strictly speaking, second and higher order terms in effects should have been included but then we could not have generalized to populations of different sizes.

The selective values $i\alpha$ and $i\beta$ of the favourable alleles have been kept constant throughout the selection process and here two important assumptions have been made. Firstly, the gene effects α and β have been defined as the difference in genotypic value between the homozygotes at the two loci as a proportion of the phenotypic standard deviation, σ. Thus, for the selective values to remain constant during selection, σ itself must remain unchanged. As selection proceeds, it would be expected that the genetic variance at other loci would decline although at the same time the environmental variance might increase as the level of homozygosity rises. We may perhaps be encouraged by the general agreement of our results with those of Latter (1965*b*) on selection effects at two additive loci within the restrictions that he imposed on the gene effects and frequencies, in that there were less assumptions made in his approach. Finally, we have taken no account of natural selection, which might be expected to alter the effective selective values of genes having correlated effects on fitness as the gene frequencies move away from their initial equilibrium values.

This work is to be continued to include more than two loci segregating simultaneously as well as non-additive gene effects. There have been several Monte Carlo studies with many loci but these have all been restricted to equal gene effects with all initial gene frequencies at one-half. Using only two loci, we have been able to analyse the interactions of the parameters at the two loci more clearly than we could have done with many loci segregating at the same time. In this restricted study, we have been able to draw attention to a situation in which linkage is likely to be important which may be of fairly general occurrence, i.e. a desirable allele in the initial population at a low frequency but with a sufficiently large effect on the character under selection that its chance of fixation is high.

Although we have succeeded in finding a reasonably simple model to explain our results, they are nevertheless a little disappointing from one point of view. Even in this simplest of all situations, we find not only curvilinearity of effects but minima in the curves. It would therefore seem rather unlikely that any general theory could be constructed to be useful in the more complex situations which must exist in practice.

A further restriction of the results, but one which can easily be removed, is that we have dealt only with populations in initial linkage equilibrium. Mather (1943)

has argued that natural selection will favour a balance between alleles at linked loci with similar effects on the character under selection, but Wright (1952) has shown that selection values have to be large and linkage very tight for such equilibrium to be maintained. In general, if loci have no epistatic effects on fitness, an unselected closed random-mating population would be expected to remain in equilibrium (Lewontin & Kojima, 1960). On the other hand, our results show that linkage disequilibrium (in the form of an excess of repulsion gametes) is likely in populations derived from crosses between selected lines or between selected lines and unselected populations. These situations need further investigation, for they have particular relevance to problems of breaking through selection limits in artificial selection.

SUMMARY

(i) A computer simulation study has been made of selection on two linked loci in small populations, where both loci were assumed to have additive effects on the character under selection with no interaction between loci. If N is the effective population size, i the intensity of selection in standard units, α and β measure the effects of the two loci on the character under selection as a proportion of the phenotypic standard deviation and c is the crossover distance between them, it was shown that the selection process can be completely specified by $Ni\alpha$, $Ni\beta$ and Nc and the initial gene frequencies and linkage disequilibrium coefficient. It is then easily possible to generalize from computer runs at only one population size. All computer runs assumed an initial population at linkage equilibrium between the two loci. Analysis of the results was greatly simplified by considering the influence of segregation at the second locus on the chance of fixation at the first (defined as the proportion of replicate lines in which the favoured allele was eventually fixed).

(ii) The effects of linkage are sufficiently described by Nc. The relationship between chance of fixation at the limit and linkage distance (expressed as $2Nc/(2Nc+1)$) was linear in the majority of computer runs.

(iii) When gene frequency changes under independent segregation were small, linkage had no effect on the advance under selection. In general, segregation at the second locus had no detectable influence on the chance of fixation at the first if the gene effects at the second were less than one-half those at the first. With larger gene effects at the second locus, the chance of fixation passed through a minimum and then rose again. For two loci to have a mutual influence on one another, their effects on the character under selection should not differ by a factor of more than two.

(iv) Under conditions of suitable relative gene effects, the influence of segregation at the second locus was very dependent on the initial frequency of the desirable allele. The chance of fixation at the first, plotted against initial frequency of the desirable allele at the second, passed through a minimum when the chance of fixation at the second locus was about 0·8.

(v) A transformation was found which made the influence of segregation at the second locus on the chance of fixation at the first almost independent of initial gene frequency at the first and of gene effects at the first locus when these are small.

(vi) In the population of gametes at final fixation, linkage was not at equilibrium and there was an excess of repulsion gametes.

(vii) The results were extended to a consideration of the effect of linkage on the limits under artificial selection. Linkage proved only to be of importance when the two loci had roughly equal effects on the character under selection. The maximum effect on the advance under selection occurred when the chance of fixation at both of the loci was between 0·7 and 0·8. When the advance under selection is most sensitive to changes in recombination value, a doubling of the latter in no case increased the advance under selection by more than about 6%. The proportion selected to give maximum advance under individual selection (0·5 under independent segregation) was increased, but only very slightly, when linkage is important.

(viii) These phenomena could be satisfactorily accounted for in terms of the time scale of the selection process and the effective size of the population within which changes of gene frequency at the locus with smaller effect must take place.

REFERENCES

Allan, J. S. & Robertson, A. (1964). The effect of initial reverse selection upon total selection response. *Genet. Res.* **5**, 68–79.

Dempster, E. R. (1955). Genetic models in relation to animal breeding problems. *Biometrics*, **11**, 525–536.

Ewens, W. J. (1963). Numerical results and diffusion approximations in a genetic process. *Biometrika*, **50**, 241–249.

Felsenstein, J. (1965). The effect of linkage on directional selection. *Genetics*, **52**, 349–363.

Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers. I. Introduction. *Aust. J. biol. Sci.* **10**, 484–491.

Gill, J. L. (1965). Effects of finite size on selection advance in simulated genetic populations. *Aust. J. biol. Sci.* **18**, 599–617.

Griffing, B. (1960). Theoretical consequences of truncation selection based on the individual phenotype. *Aust. J. biol. Sci.* **13**, 307–343.

Kimura, M. (1955). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harb. Symp. quant. Biol.* **20**, 33–55.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *Ann. math. Statist.* **28**, 882–901.

Latter, B. D. H. (1965*a*). The response to artificial selection due to autosomal genes of large effect. I. Changes in gene frequency at an additive locus. *Aust. J. biol. Sci.* **18**, 585–598.

Latter, B. D. H. (1965*b*). The response to artificial selection due to autosomal genes of large effect. II. The effects of linkage on limits to selection in finite populations. *Aust. J. biol. Sci.* **18**, 1009–1023.

Latter, B. D. H. (1966*a*). The response to artificial selection due to autosomal genes of large effect. III. The effects of linkage on the rate of advance and approach to fixation infinite populations. *Aust. J. biol. Sci.* **19**, 131–146.

Latter, B. D. H. (1966*b*). The interaction between effective population size and linkage intensity under artificial selection. *Genet. Res.* **7**, 313–323.

Lewontin, R. C. & Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**, 458–472.

Martin, F. G. & Cockerham, C. C. (1960). High speed selection studies. *Biometrical Genetics* (O. Kempthorne, ed.), pp. 35–45. Pergamon Press.

Mather, K. (1943). Polygenic inheritance and natural selection. *Biol. Rev.* **18**, 32–64.

Nei, M. (1963). Effect of selection on the components of genetic variance. *Statistical Genetics and Plant Breeding* (W. D. Hanson & H. F. Robinson, eds.), Publ. 982, National Academy of Sciences, National Research Council, Washington, D.C., pp. 501–515.

QURESHI, A. W. (1963). A Monte Carlo evaluation of the role of finite population size and linkage in response to continuous mass selection. Technical Report MC 6, Statistical Laboratory, Iowa State University.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proc. R. Soc.* B, **153**, 234–249.

WRIGHT, S. (1952). The genetics of quantitative variability. *Quantitative Inheritance* (E. C. R. Reeve & C. H. Waddington, eds.), pp. 5–41. London: H.M.S.O.

2

On the theory of artificial selection in finite populations

by

William G. Hill

Genet. Res., Camb. (1969), **13**, *pp.* 143–163
With 2 text-figures
Printed in Great Britain

On the theory of artificial selection in finite populations*

By W. G. HILL†
Statistical Laboratory, Iowa State University, Ames, Iowa, 50010, U.S.A.

(*Received* 4 *June* 1968)

1. INTRODUCTION

In a simple type of artificial selection programme individuals are ranked on their own phenotype for some quantitative trait and the highest ranking individuals are selected to be parents of the next generation. Prediction equations for changes in gene frequency with this form of selection have been derived for models in which the population is assumed to be infinitely large (Haldane, 1931; Kimura, 1958; Griffing, 1960; Latter, 1965). The truncation point, which is the value on the phenotypic scale exceeded only by selected individuals, can be assumed to be constant for specified gene frequencies and genotypic effects if the population size is infinite. However, in a finite population the truncation point must be a random variable with its value dependent on the genotypes and environmental deviations of the individuals actually present in the population. Kojima (1961) has derived formulae for expected changes in gene frequency at a single locus in finite populations, but an assumption of his model is that the effects of individual genes on the quantitative trait are small relative to the phenotypic standard deviation. Curnow & Baker (1968) have extended Kojima's results to repeated cycles of selection by using a beta distribution to approximate the distribution of gene frequencies.

In this paper a rather restricted model is analysed exactly. Predictions of changes in gene frequency are obtained for the case where there is selection on the basis of the individual phenotype (mass selection), but the quantitative trait is affected by the genotype at only one locus and by random environmental deviations. The theory is developed initially for a single cycle of selection, but is then extended to cover repeated generations of selection in a finite monecious diploid population in which there is random mating. Some of the formulae obtained are evaluated numerically for the case of normally distributed environmental deviations.

These numerical results are used to check some approximate methods which may be used to study changes in gene frequency in finite populations. These approximations involve infinite population models or assumptions of genes with small effect on the quantitative trait. In particular, some of the theory of limits to artificial selection in finite populations (Robertson, 1960; Allan & Robertson, 1964; Hill & Robertson, 1966) has been based on results of Kimura (1957) for the chance of fixation of single genes. Kimura used a haploid model and adopted a diffusion equa-

* Journal Paper No. J-6036, Iowa Agricultural and Home Economics Experiment Station, Ames, Iowa, Project No. 1669. Supported by National Institute of Health, Grant No. G.M. 13827.

† Present address: Institute of Animal Genetics, Edinburgh, 9.

tion, which is continuous in time and gene frequency. In extending these results to artificial selection programmes Robertson (1960) had to use results from infinite population theory to compute selective values of the genes affecting the metric trait.

This paper thus falls into two separate parts. In the first a mathematical theory of response to artificial selection for single loci is developed and in the second numerical checks are made to test the accuracy of more simple, approximate, formulae.

2. THEORETICAL ANALYSIS

A generation, which comprises one cycle of selection, may be considered in two successive stages. In the initial stage a sample of say, M, individuals is obtained at random from reproduction among the parents. These M individuals are a sample from a conceptual population of infinite size, comprised of all possible progeny genotypes and phenotypes from the given set of parents, with the probability distribution of genotypes among these M individuals depending on the mating system and parental genotypic frequencies. In the second stage the M individuals are ranked on phenotype and the top ranking N, say, are selected to be parents of the next generation. The second stage, namely of selection, will be discussed first as this is more difficult. Thus we consider a subpopulation of M individuals each of which has specified genotype, but not phenotypic value.

(i) *Single stage of selection from a finite sample with specified genotypes*

For simplicity let us assume that there are only two kinds of genotype, denoted A_1 and A_2. These may be regarded as either haploid individuals or the only two genotypes segregating in a backcross to a homozygous line. Extension to three or more genotypes is straightforward and will be given later.

The phenotypic values of individuals of genotype A_1, for example, are random variables because there are chance environmental effects and, in general, because of segregation at other loci, but these loci are assumed neutral in the model. Let us assume that the phenotypic values have continuous probability density functions and cumulative distribution functions given by

$$A_1: f_1(x), \quad F_1(x), \quad -\infty < x < \infty;$$
$$A_2: f_2(x), \quad F_2(x), \quad -\infty < x < \infty.$$

The mean of each distribution can be interpreted as the appropriate genotypic value, and deviations from the mean come from environmental effects.

Let us assume that in some sample of M individuals there are M_1 of type A_1 and M_2 of type A_2, with $M_1 + M_2 = M$. The N individuals with the best phenotype are selected and among these the numbers of A_1 and A_2 individuals will depend on the actual phenotypes of the M individuals. Thus we wish to compute the conditional probability of selecting $N_1 A_1$ and $N_2 = N - N_1$, A_2 individuals, conditional on M_1 and M_2 and also, of course M and N. Let this probability be denoted $p(N_1|M_1)$ where, for $N_1, N_2 \geqslant 0$, N_1 must lie in the range

$$\max(0, N - M_2) \leqslant N_1 \leqslant \min(M_1, N),$$

where max and min denote the greater and smaller terms, respectively, in their arguments. The probability $p(N_1|M_1)$ will now be derived using order statistics for mixed distributions.

Imagine that the poorest individual selected has phenotype in the range x to $x+dx$, with N_1A_1 and N_2A_2 selected. Then either the N_1th largest of the A_1's is in $[x, x+dx]$, so that N_2 of the A_2's have phenotype above $x+dx$ and $(M_2-N_2)A_2$'s have phenotype below x *or* the N_2th ranking of the A_2's is in $[x, x+dx]$ with N_1A_1's above $x+dx$ and $(M_1-N_1)A_1$'s below x. For $dx \to 0$ these events are mutually exclusive. From the theory of order statistics we know that the probability that the N_1th largest A_1 from the sample of M_1 lies in $[x, x+dx]$ is

$$\frac{M_1!}{(N_1-1)!\,(M_1-N_1)!}[F_1(x)]^{M_1-N_1}[1-F_1(x)]^{N_1-1}f_1(x)\,dx$$

and the probability that only N_2A_2's have phenotype superior to $x+dx$ is, as $dx \to 0$,

$$\frac{M_2!}{N_2!\,(M_2-N_2)!}[F_2(x)]^{M_2-N_2}[1-F_2(x)]^{N_2}.$$

These probabilities are independent. Also, summing over the two alternatives that the Nth ranking is A_1 or A_2, we obtain the probability that the Nth ranking lies in $[x, x+dx]$ with N_1A_1 and N_2A_2 selected, which is

$$\begin{aligned}
&\frac{M_1!}{(N_1-1)!\,(M_1-N_1)!}[F_1(x)]^{M_1-N_1}[1-F_1(x)]^{N_1-1}f_1(x)\,dx \\
&\qquad\qquad \times \frac{M_2!}{N_2!(M_2-N_2)!}[F_2(x)]^{M_2-N_2}[1-F_2(x)]^{N_2} \\
&+\frac{M_2!}{(N_2-1)!\,(M_2-N_2)!}[F_2(x)]^{M_2-N_2}[1-F_2(x)]^{N_2-1}f_2(x)\,dx \\
&\qquad\qquad \times \frac{M_1!}{N_1!(M_1-N_1)!}[F_1(x)]^{M_1-N_1}[1-F_1(x)]^{N_1}.
\end{aligned}$$

Integrating over x and simplifying, we obtain

$$p(N_1|M_1) = \binom{M_1}{N_1}\binom{M_2}{N_2}\int_{-\infty}^{\infty}\{[F_1(x)]^{M_1-N_1}[F_2(x)]^{M_2-N_2}[1-F_1(x)]^{N_1-1}[1-F_2(x)]^{N_2-1}$$

where
$$\times\{N_1[1-F_2(x)]f_1(x)+N_2[1-F_1(x)]f_2(x)\}\}\,dx, \qquad (1)$$

$$\max(0, N-M_2) \leqslant N_1 \leqslant \min(M_1, N) \quad \text{and} \quad M_2 = M-M_1, N_2 = N-N_1.$$

Generalization to $g > 2$ genotypes with distributions $F_1(x), \ldots, F_g(x)$ is immediate. The Nth largest individual may be from each alternative type. If there are $M_1, \ldots, M_g$ with $\sum_{i=1}^{g} M_i = M$ in the original sample of each type, the probability that $N_1, \ldots, N_g$ are selected becomes

$$\begin{aligned}
p(N_1, \ldots, N_g|M_1, \ldots, M_g) = \int_{-\infty}^{\infty}\prod_{i=1}^{g}\binom{M_i}{N_i}[F_i(x)]^{M_i-N_i}[1-F_i(x)]^{N_i} \\
\times\sum_{j=1}^{g}N_j[1-F_j(x)]^{-1}f_j(x)\,dx. \qquad (2)
\end{aligned}$$

Thus we have obtained a general equation for the distribution of the genotypes of individuals selected on the basis of phenotype from a finite population.

If the genotypes all have identical distributions, $F_1(x) = F_2(x) = \ldots = F_g(x)$, equation (2) reduces to

$$p(N_1, \ldots, N_g|M_1, \ldots, M_g) = \prod_{i=1}^{g} \binom{M_i}{N_i} \Big/ \binom{M}{N}$$

which is the hypergeometric distribution. So with neutral genes, the problem is reduced to one of random sampling of N out of M without replacement.

If there is complete dominance at a single locus having only two alleles, A and a, some simplification of the formulae for three genotypes is possible if we assume that the distributions of environmental deviations as well as genotypic values are the same for both genotypes carrying the dominant allele, A. Letting the subscripts 1, 2 and 3 refer to AA, Aa and aa individuals, respectively, we have $F_1(x) = F_2(x)$ and thus

$$p(N_1, N_2, N_3|M_1, M_2, M_3) = p(N_1+N_2|M_1+M_2) \binom{M_1}{N_1} \binom{M_2}{N_2} \Big/ \binom{M_1+M_2}{N_1+N_2}, \qquad (3)$$

where $p(N_1+N_2|M_1+M_2)$ is obtained by appropriate substitution in equation (1).

In the haploid case of equation (1) the expected frequency of A_1 among the selected individuals is, of course,

$$E(N_1/N|M_1) = \frac{1}{N} \sum_{\max(0, N-M_2)}^{\min(M_1, N)} N_1 p(N_1|M_1)$$

and in the diploid case the expected frequency of the allele A is

$$\frac{1}{N} \sum_{R} (N_1 + N_2/2)\, p(N_1, N_2, N_3|M_1, M_2, M_3), \qquad (4)$$

where R denotes all possible combinations of N_1, N_2 and N_3 such that

$$N_1+N_2+N_3 = N \quad \text{and} \quad N_1, N_2, N_3 \geqslant 0.$$

(ii) *The complete cycle of selection*

Our analysis has so far only been in terms of selection from a finite sample with specified genotypes. The distribution of these M genotypes available for selection will depend on the genotypes of their parents, the mating system, fertility differences among the parents and viability differences of the individuals prior to artificial selection. Let us consider just the case of three genotypes and assume that the probability that there are M_1, M_2 and M_3 individuals of genotype AA, Aa and aa available for selection is $\pi(M_1, M_2, M_3|S)$, where S specifies the parental genotypes, mating system, etc., and $M_1+M_2+M_3 = M$. If individual selection is practised among these M individuals, the probability $Q(N_1, N_2, N_3|S)$ of selecting N_1, N_2 and N_3 of type AA, Aa and aa, respectively, is

$$Q(N_1, N_2, N_3|S) = \sum_{C} p(N_1, N_2, N_3|M_1, M_2, M_3)\, \pi(M_1, M_2, M_3|S), \qquad (5)$$

where $p(N_1, N_2, N_3|M_1, M_2, M_3)$ is given by (2) and summation (C) is taken over all values of M_1, M_2 and M_3 such that $M_1+M_2+M_3 = M$.

In a regular breeding system in a monecious population in which each generation N individuals breed M progeny, from which the best N are selected, we can replace S in equation (5) by the numbers N_{1t}, N_{2t}, N_{3t} of genotypes at generation t, thus obtaining the transition probability $Q(N_{1,t+1}, N_{2,t+1}, N_{3,t+1}|N_{1t}, N_{2t}, N_{3t})$. Under our model this is independent of t.

(iii) *Repeated cycles of selection with random mating*

The model which will be considered in detail is where there is random mating, including random selfing, among N monecious diploid individuals in each generation, and there are no fertility or viability differences. Then the progeny are multinomially distributed, and

$$\pi(M_1, M_2, M_3|S) = \binom{M}{M_1 M_2 M_3} \left(\frac{i}{2N}\right)^{2M_1} \left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{M_2} \left(1-\frac{i}{2N}\right)^{2M_3},$$

where there are $i = 2N_{1t} + N_{2t}A$ alleles among the parents at generation t. The gene frequency is $i/2N$. Thus

$$Q(N_{1,t+1}, N_{2,t+1}, N_{3,t+1}|N_{1t}, N_{2t}, N_{3t}) = Q(N_{1,t+1}, N_{2,t+1}, N_{3,t+1}|i)$$

so that with no selection and random mating the distribution of genotypes among individuals of the next generation is a function only of i and not the genotypic frequencies. Now we can construct a transition probability matrix, $\mathbf{P}$, for changes in gene frequency from generation to generation, and can ignore the genotypic distribution of both the parental and progeny populations. We also assume that these transition probabilities are independent of t; i.e. that the distribution of genotypic values does not change with time, nor does the mating system. Let $\mathbf{P}$, with elements $(p_{ij}), i, j = 0, \ldots, 2N$ be the conditional probability that there are jA alleles among the N parents at generation $t+1$ given that there were i among the parents at generation t. Thus

$$p_{ij} = \sum_{N_1, N_2, N_3 \text{ for } 2N_1 + N_2 = j} Q(N_1, N_2, N_3|i) \quad (i, j = 0, \ldots, 2N),$$

where summation is taken over all combinations of N_1, N_2, N_3 such that there are $2N_1 + N_2 = jA$ alleles among the parents of the next generation. Combining all the relevant formulae we obtain

$$p_{ij} = \sum_{N_1, N_2, N_3 \text{ for } 2N_1 + N_2 = j} \sum_{C} \binom{M}{M_1 M_2 M_3} \left[\frac{i}{2N}\right]^{2M_1} \left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{M_2} \left[1-\frac{i}{2N}\right]^{2M_3}$$
$$\times \int_{-\infty}^{\infty} \prod_{h=1}^{3} \binom{M_h}{N_h} [F_h(x)]^{M_h - N_h} [1 - F_h(x)]^{N_h} \sum_{k=1}^{3} N_k [1 - F_k(x)]^{-1} f_k(x)\, dx. \quad (6)$$

Changes in the distribution of gene frequency for several cycles of selection can be obtained by repeated multiplication of the matrix $\mathbf{P}$.

3. NUMERICAL EVALUATION OF THE FORMULAE FOR NORMALLY DISTRIBUTED ENVIRONMENTAL DEVIATIONS

Let us assume that environmental deviations are normally distributed about the genotypic value, an assumption which is made in most theoretical predictions of selection advance. The integrals of equations (1) and (2) cannot then be evaluated without recourse to numerical methods unless, of course, $F_1(x) = F_2(x) = F_3(x)$ for the two allele diploid model we shall investigate. For the case of additive gene action on the quantitative trait let the phenotypic values of AA individuals have a normal distribution with mean $\mu + \alpha\sigma$ and variance σ^2, i.e. have the $N(\mu + \alpha\sigma, \sigma^2)$ distribution and similarly let $N(\mu + \alpha\sigma/2, \sigma^2)$ and $N(\mu, \sigma^2)$ be the distributions of Aa and aa individuals, respectively, where $-\infty < \mu < \infty$, $-\infty < \alpha < \infty$ and $0 < \sigma^2 < \infty$. But $p(N_1, N_2, N_3|M_1, M_2, M_3)$ is dependent only on α in this model so that equation (2) can be evaluated using the normal distributions $N(\alpha/2, 1)$, $N(0, 1)$ and $N(-\alpha/2, 1)$ for AA, Aa and aa individuals. Thus α is the difference between the phenotypic values of the two homozygotes as a proportion of the environmental standard deviation. Letting $\phi(x)$ and $\Phi(x)$ denote the density and distribution functions of the standardized normal distribution, $N(0, 1)$, equation (2) for the additive model becomes

$$p(N_1, N_2, N_3|M_1, M_2, M_3) = \int_{-\infty}^{\infty} \prod_{i=1}^{3} \binom{M_i}{N_i} [\Phi(x - \alpha + \tfrac{1}{2}i\alpha)]^{M_i - N_i} [\Phi(-x + \alpha - \tfrac{1}{2}i\alpha)]^{N_i} \times \sum_{j=1}^{3} \{N_j [\Phi(-x + \alpha - \tfrac{1}{2}j\alpha)]^{-1} \phi(x - \alpha + \tfrac{1}{2}j\alpha)\} dx \quad (7)$$

since, by symmetry, $1 - \Phi(x) = \Phi(-x)$.

With a model of complete dominance the alternative homozygotes have also been assumed to differ by $\alpha\sigma$ units in genotypic value and to have normally distributed phenotypic values. Thus from equations (1) and (3) we have

$$p(N_1, N_2, N_3|M_1, M_2, M_3) = \binom{M_1}{N_1}\binom{M_2}{N_2}\binom{M_3}{N_3} \int_{-\infty}^{\infty} [\Phi(x - \tfrac{1}{2}\alpha)]^{M_1 + M_2 - N_1 - N_3} [\Phi(x + \tfrac{1}{2}\alpha)]^{M_3 - N_3} \times [\Phi(-x + \tfrac{1}{2}\alpha)]^{N_1 + N_2 - 1} [\Phi(-x - \tfrac{1}{2}\alpha)]^{N_3 - 1} \times [(N_1 + N_2)\, \Phi(-x - \tfrac{1}{2}\alpha)\, \phi(x - \tfrac{1}{2}\alpha)\, dx + N_3\, \Phi(-x + \tfrac{1}{2}\alpha)\, \phi(x + \tfrac{1}{2}\alpha)\, dx]. \quad (8)$$

The probabilities $p(N_1, N_2, N_3|M_1, M_2, M_3)$ are much less quickly computed for all N_1, N_2, N_3 with equation (7) than equation (8). In the latter numerical integration need only be performed for the range of possible values of $N_1 + N_2$.

Equations (7) and (8) were integrated by Simpson's rule over the region

$$-5{\cdot}12 \leqslant x \leqslant 5{\cdot}12$$

using an I.B.M. 360/50 computer with double-precision arithmetic. The values of $\Phi(x)$ were previously tabulated in the computer in the same way. Since

$$p(N_1, N_2, N_3|M_1, M_2, M_3)$$

is a probability mass function it must sum to unity over the range of N_1, N_2 and N_3

possible for specified M_1, M_2 and M_3. The step length for integration was taken sufficiently small that

$$|\Sigma p(N_1, N_2, N_3|M_1, M_2, M_3) - 1| < 10^{-7}.$$

The range $-5{\cdot}12 \leqslant x \leqslant 5{\cdot}12$ was found to be adequately wide, since quantities like $[\Phi(x)]^{M-N}[\Phi(x)]^N \phi(x)$ are very small unless x is close to zero.

In Table 1 some examples of the form of $p(N_1, N_2, N_3|M_1, M_2, M_3)$ are given for the case of additive gene action. The expectations of the genotypic and gene frequencies among the selected individuals are also shown. In the next section we shall use the exact results obtained by numerical integration to check various approximate formulae for selection advance in both single and repeated cycles of selection.

Table 1. *Probabilities of selecting each possible combination of genotypes in a single stage of selection for an additive gene*

($M_1 = 4$, $M_2 = 8$, $M_3 = 4$ and $N = 4$.)

			$p(N_1, N_2, N_3\|M_1, M_2, M_3)$		
N_1	N_2	N_3	$\alpha = 0$	$\alpha = 0{\cdot}2$	$\alpha = 0{\cdot}8$
0	0	4	0·000549	0·000283	0·000028
0	1	3	0·017582	0·010634	0·001775
0	2	2	0·092308	0·065634	0·018227
0	3	1	0·123077	0·102878	0·047574
0	4	0	0·038462	0·037795	0·029133
1	0	3	0·008791	0·006232	0·001657
1	1	2	0·105494	0·087919	0·038877
1	2	1	0·246154	0·241161	0·177522
1	3	0	0·123077	0·141750	0·173873
2	0	2	0·019780	0·019321	0·013590
2	1	1	0·105494	0·121134	0·141784
2	2	0	0·092308	0·124597	0·242910
3	0	1	0·008791	0·011830	0·021990
3	1	0	0·017582	0·027813	0·086063
4	0	0	0·000549	0·001019	0·004994
$E(N_1/N)$			0·250000	0·282543	0·383159
$E(N_2/N)$			0·500000	0·498837	0·481673
$E(N_3/N)$			0·250000	0·218620	0·135168
$E[(N_1+N_2/2)/N]$			0·500000	0·531961	0·623996

4. COMPARISON OF RESULTS FROM EXACT AND APPROXIMATE METHODS

(i) *Single stage of selection from a sample with specified genotypes*

If there are M_1, M_2 and M_3 individuals of genotype AA, Aa and aa respectively from which selection is made, the expected gene frequency in selected individuals is given by (4). However, as $M \to \infty$ the average gene frequency among selected individuals is readily computed, for the trunaction point T is no longer a random variable. With additive gene action T must satisfy the following equation on the standardized scale

$$M_1 \Phi(-T + \tfrac{1}{2}\alpha) + M_2 \Phi(-T) + M_3 \Phi(-T - \tfrac{1}{2}\alpha) = N \qquad (9)$$

as $M \to \infty$ where, for example $\Phi(-T + \frac{1}{2}\alpha) = 1 - \Phi(T - \frac{1}{2}\alpha)$ is the proportion of AA

individuals which have phenotype superior to T and are selected. The mean frequency q', of A alleles among the selected individuals is then

$$q' = \frac{M_1}{N}\Phi(-T+\tfrac{1}{2}\alpha)+\frac{M_2}{2N}\Phi(-T). \tag{10}$$

Equation (10) is easily evaluated and requires little computation.

Table 2. *Expected change in gene frequency when selecting N individuals from a population of size M in exact Hardy–Weinberg frequencies*

(The change in gene frequency is tabulated for infinite M, and changes at other values of M as a percentage difference, $P = [(q'-q)_m/(q'-q)_\infty - 1]\times 100$.)

		Additive				Complete dominant			
	M ...	8	16	32	$\to\infty$	8	16	32	$\to\infty$
q	α		P		$q'-q$		P		$q'-q$
					(1) $N/M = \frac{1}{4}$				
0·25	0·2	—	0·91	0·36	0·024192	—	0·83	0·33	0·035937
	0·8	—	1·17	0·48	0·098819	—	0·74	0·28	0·140654
0·5	0·2	2·30	0·78	0·30	0·031713	1·65	0·51	0·18	0·030604
	0·8	2·23	0·73	0·27	0·123099	−0·27	−0·32	−0·21	0·104206
0·75	0·2	—	0·65	0·24	0·023391	—	0·33	0·09	0·011176
	0·8	—	0·29	0·07	0·086674	—	0·61	−0·35	0·034964
					(2) $\alpha = 0{\cdot}4$				
q	N/M								
0·25	$\frac{1}{2}$	—	1·58	0·72	0·029723	—	1·62	0·74	0·044569
	$\frac{1}{8}$	—	−0·84	−0·60	0·064769	—	−1·20	−0·75	0·093064
	$\frac{1}{16}$	—	—	−2·40	0·078694	—	—	−2·51	0·110752
0·5	$\frac{1}{2}$	3·81	1·59	0·72	0·039630	3·71	1·55	0·71	0·039431
	$\frac{1}{8}$	—	−1·18	−0·75	0·081449	—	−1·74	−0·89	0·071782
	$\frac{1}{16}$	—	—	−2·45	0·096986	—	—	−2·43	0·081990
0·75	$\frac{1}{2}$	—	1·58	0·72	0·029723	—	1·41	0·64	0·014637
	$\frac{1}{8}$	—	−1·45	−0·87	0·057791	—	−1·88	−1·05	0·024820
	$\frac{1}{16}$	—	—	−2·48	0·067622	—	—	−2·31	0·027830

In Table 2 predictions of expected change in gene frequency from a single stage of selection using the finite population and infinite population methods are compared for both additive and completely dominant gene action. The configurations of genotypic frequency among the M individuals are chosen such that $M_2^2 = 4M_1M_3$, with the original frequency q being $q = (M_1+M_2/2)/M$. Thus for $q = 0{\cdot}25$ and $M = 32$ we have $M_1 = 2, M_2 = 12$ and $M_3 = 18$. These genotypic frequencies are those corresponding to the Hardy–Weinberg equilibrium frequencies, but since we are only considering one sample they may be assumed to have occurred by chance. Other possible configurances have not been considered separately. In Table 2 the predicted changes in gene frequency $(q'-q)$ computed with the infinite population model (equation (10) for additive gene action) are given, and the expected changes using the finite model expressed as a proportion of these. The results of the table

indicate that the infinite population model gives a very close prediction of the response expected from finite populations. Even when as few as 2 individuals out of 16 are chosen the error scarcely exceeds 2 % of the mean change in gene frequency.

(ii) *Complete cycle of selection with random mating*

In a complete generation or cycle of selection there is sampling of progeny followed by selection of parents for the next generation. We shall only consider the case where the genotypes among the M progeny are multinomially distributed with expected frequencies q^2, $2q(1-q)$ and $(1-q)^2$ for AA, Aa and aa individuals, where q is the frequency of A among the parents. General formulae for this model, assuming a random mating monecious population, have been given in an earlier section. The expected gene frequency $E(q')$ among the parents of the next generation is, for complete dominance and integral $2Nq$,

$$E(q') = E(j/2N|q = i/2N) = \frac{1}{2N}\sum_{j=0}^{2N} jp_{ij} \tag{11}$$

where p_{ij} is given in equation (6). The model has been restricted by assuming that there are N parents in each of the two generations, but relaxation of this assumption is straightforward. Also integration has been carried out only for the case of complete dominance so that equation (8) could be used to reduce computation time.

An approximate method for obtaining $E(q')$ has been given by Kojima (1961). He showed that for small values of α (the gene effect in standard deviations) such that α^2, α^3, etc. could be ignored relative to α, the mean change in gene frequency

$$\left.\begin{aligned} &\delta q = E(q') - q \\ \text{is} \quad &\delta q \sim k\alpha q(1-q)^2 \end{aligned}\right\} \tag{12}$$

for complete dominance. Kojima calls k a 'generalized selection differential', and Pike (1969) has shown that if the phenotypic values are normally distributed about the genotypic values k becomes the mean of the highest N order statistics in a sample of size M from a single standardized normal distribution.

As M becomes infinitely large the value of k can be obtained directly from tables of the normal distribution, and may be denoted i, the standardized selection differential. Thus $\lim_{M\to\infty} k = i$ for N/M constant. Equation (12) is then the well known approximate formula for the change in gene frequency with truncation selection (Haldane, 1931; Kimura, 1958; Griffing, 1960; Latter, 1965) in which $i\alpha$ is the selective value of the allele A. Latter (1965) has studied the errors associated with this approximation for predicting changes in gene frequency in infinite population. The exact values for q' in infinite population can, of course, be obtained from (10).

In Table 3 the approximate and exact methods are compared for a choice of values of the parameters α, N/M, q and M. Predictions of change of gene frequency, δq, have been computed for the exact method (equation (11)) and are tabulated as a proportion of the change predicted by the simple form $\delta q = k\alpha q(1-q)^2$. The values of k were obtained from tables of the expectations of order statistics from the normal distribution, which are given to 10 decimal places by Teichroew (1956). In the

limiting case of $N \to \infty$, k equals i in the approximate formulae, and in the exact formulation the finite population predictions are replaced by the exact infinite predictions of equation (10), where M_1, M_2 can be replaced by Mq^2 and $2Mq(1-q)$. The values for $N \to \infty$ are thus tests of the infinite model approximations, but at the same time serve as limiting values for the finite model approximations of Kojima in which it is assumed that gene effects (α) are small.

Table 3. *Response from one full cycle of selection for complete dominance*

(M progeny are taken at random from parents in Hardy–Weinberg equilibrium with gene frequency q, and N are selected. The response (δq) is tabulated as a percentage deviation from $k\alpha q(1-q)^2$, i.e. as $[\delta q/k\alpha q(1-q)^2 - 1] \times 100$, where k is the mean of first N from M order statistics from the standardized normal distribution.)

		M			
q	α	4	8	16	$\to \infty$
		$N/M = \frac{1}{2}$			
0·25	0·2	−0·27	−0·23	−0·20	−0·18
	0·8	−4·23	−3·60	−3·20	−2·73
0·5	0·2	−0·32	−0·31	−0·30	−0·29
	0·8	−4·87	−4·73	−4·66	−4·58
0·75	0·2	−0·41	−0·47	−0·51	−0·55
	0·8	−6·27	−7·12	−7·64	−8·22
		$\alpha = 0{\cdot}4$			
q	N/M		8	16	$\to \infty$
0·25	$\frac{1}{4}$	—	+0·09	+0·24	+0·42
	$\frac{1}{8}$	—	—	+0·32	+0·84
0·5	$\frac{1}{4}$	—	−7·24	−7·60	−7·99
	$\frac{1}{8}$	—	—	−11·91	−12·50
0·75	$\frac{1}{4}$	—	−11·48	−12·16	−21·90
	$\frac{1}{8}$	—	—	−18·22	−19·32

We see in Table 3 that the approximation for the finite model is rarely much poorer than with an infinite population. Since essentially the same assumptions about the size of α are made in each case, we should not be surprised to observe that a poor fit between the predictions is only found with finite populations for values of α and selection intensity (i.e. N/M) which lead to inadequate approximation in infinite population.

Kojima (1961) also derived formulae for the variance of change in gene frequency based on the same assumptions as the mean change. For complete dominance this is

$$V(\delta q) \sim \frac{q(1-q)}{2N}[1 + k\alpha(1-q)(1-3q)]. \tag{13}$$

Some checks on the accuracy of (13) have been made against the exact finite population prediction, obtained by finding $E(q')^2$ by extension of equation (11). Again, as we would expect the approximate and exact methods agree well except at the highest values of $\alpha(0{\cdot}8)$ and selection intensity.

Attention should perhaps be drawn to the fact that when we calculated the expected change in gene frequency from a population with specified numbers M_1, M_2, M_3 of each genotype a good approximation was obtained using an infinite population prediction (10). For small α equation (10) reduces to $q' = q + i\alpha q(1-q)^2$ for complete dominance, where $q = (2M_1 + M_2)/2M$ and i is the infinite population standardized selection differential. However, when the M individuals in the population are themselves a sample of genotypes then the selection differential should be calculated as k from order statistics for the appropriate *finite* population size. In the latter case we obtain a reasonable fit using order statistics from the normal distribution because the combined (binomial) distribution of genotypic values and (normal) distribution of environmental values is close to normal in form.

(iii) *Chance of fixation of single genes*

The theory of limits in artificial selection in finite populations developed by Robertson (1960) is based on the concept of the chance of fixation, $u(q_0)$, which is the probability that an allele with initial frequency q_0 will eventually be fixed in the population. Using a diffusion equation (the Kolmogorov backward equation), Kimura (1957, 1962) showed that, for example, the chance of fixation of a dominant allele with selective value was

$$u(q_0) = \int_0^{q_0} e^{Ns(1-x)^2}\,dx \Big/ \int_0^1 e^{Ns(1-x)^2}\,dx. \tag{14}$$

To describe the response to artificial selection the selective value has been taken as $s = i\alpha$ (Robertson, 1960; Hill & Robertson, 1966). The model used in (14) is continuous and haploid in form, so it seemed necessary to check the accuracy of (14) for diploids with artificial selection and discrete generations. Previously Ewens (1963) has made numerical tests on the errors resulting from use of the diffusion equation, but only for haploid individuals and additive gene action.

The approximate results were obtained by numerical integration of (14), using Simpson's rule, where s was replaced by $i\alpha$ and also by $k\alpha$, with k computed for a few pairs of N and M values.

Exact results for our model of a diploid monecious random mating population with stationary transition probabilities were obtained from the matrix $\mathbf{P}$ (equation (6)). A vector $\mathbf{v}_{(0)}$ with elements $v_{i(0)}$ was first constructed, where $v_{i(0)} = i/2N$, $i = 0, \ldots, 2N$. Then successive products $\mathbf{v}_{(1)} = \mathbf{P}\mathbf{v}_{(0)}, \mathbf{v}_{(2)} = \mathbf{P}\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(t)} = \mathbf{P}\mathbf{v}_{(t-1)}$ were computed. An element $v_{i(t)}$ is therefore the expected gene frequency at time t for an initial frequency of $i/2N$. Iteration was continued for at least $6N$ generations so that the ratio of changes in gene frequency in successive generations

$$(v_{i(t)} - v_{i(t-1)})/(v_{i(t-1)} - v_{i(t-2)})$$

became sufficiently constant that the chance of fixation, $\lim_{t\to\infty} v_{i(t)}$, could be predicted to 5 decimal places by fitting an exponential curve to the last 3 values of $v_{i(t)}$ by the δ^2 method (Aitken, 1926). This iterative method of obtaining the chance of fixation was preferred to more direct methods, since expected gene frequencies at inter-

mediate generations were required for further tests on approximate methods which will be described in the next section.

In Tables 4 and 5 comparisons are shown of the chance of fixation computed by the exact method and by the diffusion approximation. Results are given for different values of $N\alpha$. Positive values imply that the dominant allele is favoured by selection,

Table 4. *Chance of fixation* $u(q_0) \times 10^4$ *for a dominant gene with truncation selection computed exactly by matrix iteration and by diffusion approximation*

(The selective value for the diffusion equation is $k\alpha$, with k computed from order statistics for specified values of N. The chance of fixation is tabulated for the diffusion results with $N \to \infty$, D_∞, others by difference from D_∞.)

		Matrix N				Diffusion N		
$N\alpha$	q_0	2	4	10	∞	10	4	2
		$[D_\infty - u(q_0)] \times 10^4$			$D_\infty \times 10^4$	$[D_\infty - u(q_0)] \times 10^4$		
				$N/M = 0{\cdot}5$				
0·2	0·25	54	—	13	2864	15	—	64
	0·5	66	—	15	5403	15	—	68
	0·75	48	—	11	7748	10	—	42
0·4	0·25	114	62	26	3254	20	73	134
	0·5	130	70	29	5812	31	75	138
	0·75	90	48	20	7990	18	43	81
0·8	0·25	252	136	57	4098	66	158	292
	0·5	245	131	55	6621	60	145	270
	0·75	159	83	34	8447	33	80	149
1·6	0·25	560	287	118	5853	130	312	583
	0·5	396	202	81	8043	94	229	437
	0·75	224	110	43	9180	45	112	216
3·2	0·25	—	436	159	8430	134	337	—
	0·5	—	156	55	9556	60	155	—
	0·75	—	66	22	9850	23	61	—
6·4	0·25	—	—	80	9837	31	—	—
	0·5	—	—	6	9989	4	—	—
	0·75	—	—	1	9998	1	—	—
−0·2	0·5	−68	—	−15	4607	−15	—	−65
−0·4	0·5	−167	−73	−30	4229	−28	−68	−126
−0·8	0·5	−271	−142	−58	3531	−50	−121	−226
−1·6	0·5	−517	−255	−101	2404	−73	−178	−340
−3·2	0·5	—	−381	−135	1088	−68	−168	—
−6·4	0·5	—	—	−119	0238	−29	—	—
				$N/M = 0{\cdot}25$				
0·8	0·25	216	117	—	5148	—	156	296
	0·5	225	118	—	7510	—	124	238
	0·75	155	79	—	8916	—	64	123
1·6	0·25	307	160	—	7595	—	220	430
	0·5	220	114	—	9146	—	122	243
	0·75	134	65	—	9684	—	52	105
−0·8	0·5	−204	−103	—	2819	—	−97	−187
−1·6	0·5	−265	−125	—	1501	—	−109	−214

negative values that the recessive allele is favoured. The continuous model with $s = i\alpha$ can be regarded as the limiting case as $N \to \infty$, with $N\alpha$ remaining constant. We find in the tables that there is mostly quite good agreement between the exact and approximate predictions of chance of fixation. As we must expect, the fit is poorest at low values of N and high values of α, for given $N\alpha$, especially when the initial frequency of the favoured allele is low (Table 5). The diffusion approximation method with $s = i\alpha$ generally overestimates the total change in expected gene frequency, $|u(q_0) - q_0|$. Thus when s is replaced by $k\alpha$ for the appropriate M and N values a better fit is obtained since $k < i$; but, except for the smallest N values, this correction may not be thought worthwhile. The values of N ($\leqslant 10$) used in this study are less than in many animal selection experiments or programmes so, in practice, k and i may differ by very little.

Table 5. *Chance of fixation* $\times 10^4$

(Computed from exact transition matrix (TM) with $N = 10$ and $M = 20$, and by diffusion approximation (DA) with selective value computed from order statistics. Diffusion result is shown as difference $D = DA - TM$.)

q_0 ...	0·05		0·1		0·5		0·9		0·95	
$N\alpha$	TM	D	TM	D	TM	D	TM	D	TM	D
0·4	720	−2	1398	−3	5783	−2	9196	+1	9598	+1
1·6	1697	+4	3046	+2	7962	−13	9668	+1	9835	0
6·4	5472	+455	7911	+338	9983	+2	9999	0	9999	0
−0·4	336	+1	692	+2	4259	−2	8799	−4	9399	−2
−1·6	86	+2	197	+2	2505	−28	8226	−28	9109	−16
−6·4	0	0	1	0	349	−82	6187	−241	8384	−138

(iv) *Rate of selection advance with repeated cycles of selection; simple transition probability matrices*

A further consequence of the diffusion approximation to the selection process for single genes in finite populations is that the distribution of gene frequencies among replicate lines, and therefore the mean gene frequency also, is a function of only Ns and the initial frequency, provided that time is measured on a scale proportional to N. This result was pointed out by Robertson (1960) and it leads to a considerable simplification of the description of the rate of advance. The chance of fixation (as $t \to \infty$) is then a function of only Ns and q_0, and we see in Table 4 that this still holds reasonably well when we compare the exact values for the chance of fixation computed for the same $N\alpha$ and N/M, but different N.

As a measure of the rate of advance we shall use the 'half-life' of the change in gene frequency, which is the time taken for the mean gene frequency to get half way from its initial to its limiting value (Robertson, 1960). Half-lives were calculated by linear interpolation between the two successive generations which had mean gene frequency spanning the half-way frequency and have been expressed proportional to the parental population size, N, in the relevant tables.

If M and N become large an excessive amount of numerical integration is re-

quired in order to evaluate the matrix **P** (equation (6)), and it becomes very difficult to carry out the computation of **P** with sufficient accuracy. Therefore it seems desirable to have a more efficient, if approximate, method for computing intermediate gene frequencies and half-lives. A simple type of transition matrix was constructed and compared with the exact matrix **P** for artificial selection in a monecious random mating population. In this simple matrix it is assumed that the gene frequency among the parents of the next generation is binomially distributed with mean $q + sq(1-q)^2$ from (12). Let us denote this matrix **B**, with elements (b_{ij}), $i,j = 0, \ldots, 2N$, which define the same transition probabilities as do the elements of **P**. Thus b_{ij} is the (approximate) probability that there are jA alleles among the N parents at generation $t+1$, given that there were i at generation t. **B** is assumed independent of t. The elements of **B** are obtained by adopting a haploid type of model. We assume that the expected gene frequency in generation $t+1$ is

$$\frac{i}{2N} + \frac{si}{2N}\left(1 - \frac{i}{2N}\right)^2$$

for complete dominance. The selective value s can be replaced by $k\alpha$ in Kojima's (1961) formulation, as we have seen in equation (12). The $2N$ alleles among the parents of the next generation are then obtained by sampling from the binomial distribution. Thus

$$b_{ij} = \binom{2N}{j}\left[\frac{i}{2N} + \frac{si}{2N}\left(1 - \frac{i}{2N}\right)^2\right]^j \left[1 - \frac{i}{2N} - \frac{si}{2N}\left(1 - \frac{i}{2N}\right)^2\right]^{2N-j}. \qquad (15)$$

Expected gene frequencies in the intermediate stages of selection and chances of fixation were obtained by repeated iteration of the matrix **B** in the same manner as described for the matrix **P**.

In Tables 6 and 7 comparison is made of the chances of fixation and half-lives, respectively, computed using matrices **P** and **B**. In **B** the selective value s is set equal to $k\alpha$ for the appropriate value of N. We find a rather better fit in Table 6 between the pairs of matrix results than we observed between the results from the diffusion and the exact method in Table 4. For the half-lives the agreement between results for different values of N and constant $N\alpha$ improves as N increases with either method. Also, for large N and small α the approximate and exact methods agree more closely with each other. This pattern of results could be predicted to some extent because the continuous model assumptions are less severely violated at large N, ignoring terms in α^2, α^3, ... becomes less serious for small α, and because the change of k with N is smaller as N becomes larger. In Tables 6 and 7 effects of population size on the genetic sampling process and on selection intensity are confounded since the parameter k is used. For constant values of Ns, but differing N, half-lives have been computed using the simplified matrix **B** and a few results are given in Table 8. Again, although some wide discrepancies occur at the higher Ns value, there is probably sufficient agreement for practical purposes because approximate values of half-lives (or other measures of rate of advance) are all we are likely to need when planning or interpreting selection experiments. Also, it should be pointed out that the apparently large discrepancies between predicted half-lives (Table 7) using

Table 6. *Chance of fixation* $\times 10^4$

(Computed by exact transition probability matrix method (P) and approximate matrix method (B) with $N/M = 0{\cdot}5$. Results are shown as deviations $P_N - P_{10}$ or $B_N - P_{10}$ from exact method with $N = 10$.)

	N ...	2		4		10	
$N\alpha$	q_0	B	P	B	P	B	P
0·4	0·25	−103	−88	−46	−36	−4	3228
	0·5	−119	−101	−53	−42	−5	5783
	0·75	−81	−70	−34	−28	−3	7970
1·6	0·25	−588	−442	−295	−169	−65	5735
	0·5	−480	−315	−234	−121	−55	7962
	0·75	−268	−181	−124	−67	−27	9137
6·4	0·25	—	—	—	—	−33	9757
	0·5	—	—	—	—	−8	9983
	0·75	—	—	—	—	−3	9997
−0·4	0·5	+86	+107	+31	+42	−4	4259
−1·6	0·5	+124	+416	+15	+154	−54	2505
−6·4	0·5	—	—	—	—	−114	349

Table 7. *Half-lives* ($\times 1000/N$ *generations*)

(Computed by exact transition probability matrix method (P) and approximate matrix method (B) with $N/M = 0{\cdot}5$. Results are shown as deviations $P_N - P_{10}$ or $B_N - B_{10}$ from exact method with $N = 10$.)

	N ...	2		4		10	
$N\alpha$	q_0	B	P	B	P	B	P
0·4	0·25	−197	−156	−76	−57	−9	1231
	0·5	−173	−131	−74	−51	−10	1412
	0·75	−147	−94	−62	−37	−10	1701
1·6	0·25	−241	−92	−115	−45	−30	1331
	0·5	−235	−69	−113	−34	−30	1414
	0·75	−199	−5	−102	−12	−34	1644
6·4	0·25	—	—	—	—	−16	645
	0·5	—	—	—	—	−22	607
	0·75	—	—	—	—	−40	762
−0·4	0·5	−84	−122	−30	−48	+8	1272
−1·6	0·5	+7	−61	+16	−24	+18	994
−6·4	0·5	—	—	—	—	−8	400

Table 8. *Half-lives* $\times 1000/N$ *generations computed with different* N *and constant* Ns *using the simplified transition matrix* B

	q_0 ...	0·25		0·5		0·75	
	N ...	8	32	8	32	8	32
Ns	1	1239	1349	1398	1456	1639	1691
	4	756	791	711	757	857	900
	−1	794	820	1076	1090	1443	1446
	−4	250	278	476	482	853	835

matrices **P** and **B** may have no practical significance for small N values. For example, with $N = 2$ half-lives of $1{\cdot}345N$ and $1{\cdot}179N$ ($N\alpha = 1{\cdot}6, q_0 = 0{\cdot}5$) both imply simply that the mean gene frequency at generation 2 is less than half-way to its expected limit and that at generation 3 more than half-way. It is possible to construct matrices other than **B** which give better approximations to the exact results. For example, diploid selection can be included in terms of selective values, and still not require numerical integration. However, the extra computation involved in setting up such matrices does not seem justified by the small increase in precision obtained. Curnow & Baker (1968) have developed an alternative method of predicting the selection advance by approximating the gene frequency distribution by a beta distribution. The accuracy of this method has recently been checked by Pike (1969) and found to be satisfactory for all but the smallest population size (4) checked.

(v) *Optimum intensity of artificial selection*

The optimum intensity of selection in an artificial selection programme has been discussed by Dempster (1955) and Robertson (1960), who pointed out that, for fixed M, the selection limit would be maximized if $N/M = 0{\cdot}5$. This conclusion is based on the diffusion equation model and assumes that N is very large so that the limit is a function of Ni. For the normal distribution $i = z/(N/M)$ where z is the ordinate of the standardized normal distribution at the truncation point. Thus $Ni = Mz$ so Ni is maximized when z is maximized at $N/M = 0{\cdot}5$. However, even in finite populations, it will now be shown that sufficient conditions for Nk to be mazimized when $N/M = 0{\cdot}5$ are for the distribution of phenotypic values to be unimodal and symmetric. Let $x_1, \ldots, x_M$ be the expected values of the order statistics as deviations from the mean of a symmetric distribution, then

$$\sum_{i=1}^{N} x_i + \sum_{i=N+1}^{M} x_i = 0$$

and by symmetry, $x_i = -x_{M-i+1}$. Substituting, we obtain

$$\sum_{i=1}^{N} x_i = \sum_{i=N+1}^{M} x_{M-i+1}$$

$$= \sum_{i=1}^{M-N} x_i$$

or

$$Nk_N = (M-N)\,k_{M-N},$$

where k_N and k_{M-N} are the means of the best N and $M-N$, respectively, ordered individuals. Therefore, as long as the approximation from the diffusion equation that Nk is a sufficient parameter holds fairly well we expect the limit to be maximized when half the population is selected and to be symmetric about this proportion.

Some checks on this prediction were made using the exact model with $M = 16$, $q_0 = 0{\cdot}5$ and $\alpha = 0{\cdot}4$ or $-0{\cdot}4$, with the limit computed for $N = 2, 4, 6, 8, 10, 12$ and 14 individuals selected each generation. Results are shown in Fig. 1, where we

observe that the curve of chance of fixation against N departs very little from symmetry. Of course the rates of approach to the limit differ widely. This is illustrated in Fig. 2, in which the mean gene frequency is plotted against number of generations (t) for $M = 16$, $q_0 = 0{\cdot}5$ and $\alpha = 0{\cdot}2$. Also in Fig. 2 we find that the time scale fits

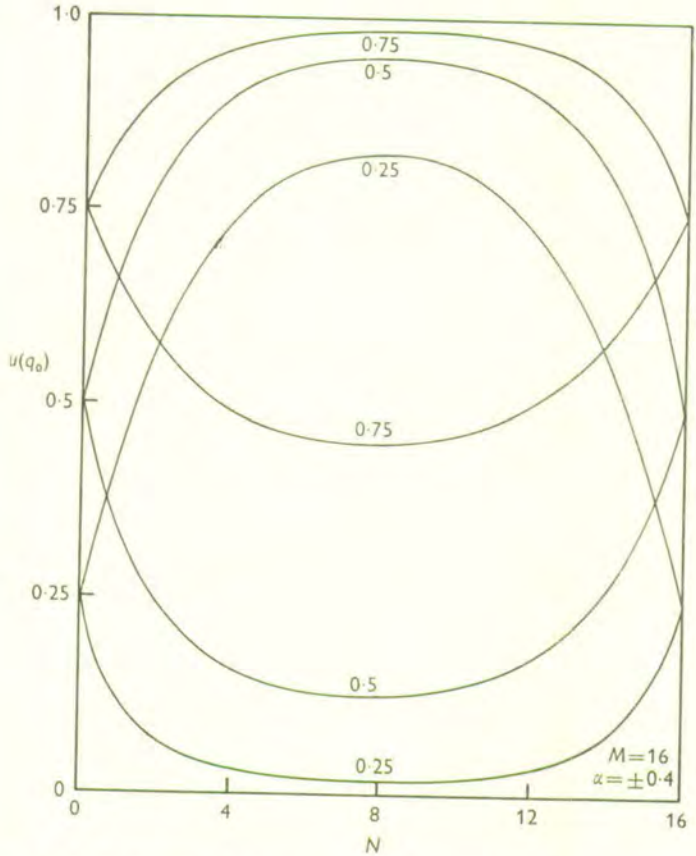

Fig. 1. The effect of selection intensity on the selection limit. The chance of fixation $u(q_0)$ is computed by the exact method and plotted for $q_0 = 0{\cdot}25$, $0{\cdot}5$ and $0{\cdot}75$, for different numbers of individuals (N) selected from 16 recorded every generation.

well with the diffusion theory in that it is inversely proportional to N. Thus we expect the same mean frequency after cN generations with population size N as with $c(M-N)$ generations with population size $M-N$, where c is a positive constant. In the example of Fig. 2 let us compare $N = 4$ with $N = 12$, where gene frequencies for some values of c are as follows:

Population size c ...	0·5	1·0	2·0	4·0	$\to \infty$
			Mean gene frequency		
4	0·55348	0·59337	0·64871	0·70424	0·72718
12	0·55366	0·59365	0·64919	0·70530	0·74129

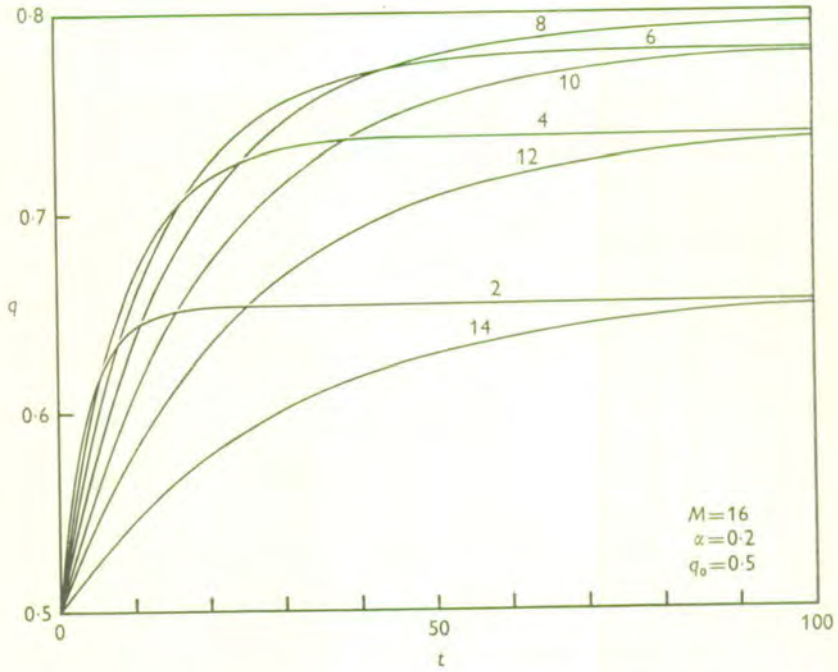

Fig. 2. The effect of selection intensity on the rate of selection advance. The mean gene frequency, q, is computed by the exact method and plotted against the number of generations (t) of selection for different numbers of individuals selected from 16 recorded every generation.

5. DISCUSSION

The formulae developed for the simple model of artificial selection in finite populations have enabled us to make checks on some of the simplifying assumptions in the theory of selection limits. The population sizes which have been tested (usually $N \leqslant 10$) are smaller than would normally be encountered in breeding schemes in which much emphasis is placed on selection within lines. Thus, in practice, we would expect population sizes to be larger and the diffusion approximation to fit better than in the examples given here. Therefore, in view of our results, we can probably conclude that the diffusion equation gives an adequate approximation for the model of a single gene in a random mating monecious population which is analysed in this paper.

At the same time, this single locus model is unlikely ever to be realized for a quantitative trait in nature, nor are monecious populations of direct interest in livestock improvement. The limitations of this approach therefore appear to rest mostly on the model adopted. However, this study should be viewed as an initial attempt to test the adequacy of some simple theory for describing artificial selection in finite populations.

A comparison of monecious and diecious models has been made by Hill & Robertson (1968) for the case of natural selection acting on viability differences at a single locus with complete dominance or heterozygote advantage. The populations

comprised 10 parents in the monecious model, and 5 of each sex in the diecious case. The alternative models led to similar results for the change in fitness (Hill & Robertson, 1968) and mean gene frequency (unpublished), and it was considered that the monecious model was adequate for descriptive purposes.

Even if there are many loci segregating for the quantitative trait the algebraic theory developed in the first part of this paper can, in principle, still be used for single generations of selection. One alternative approach is merely to evaluate the probability of selection of each possible genotype. As in the single locus situation we have discussed, the only variation of phenotypic value about genotypic value would be attributed to environmental deviation. The expected change of gene frequency at a single locus can then be obtained by summation of selection probabilities over all possible genotypes. Alternatively, segregation at other loci can be included as variation in the phenotypic values about the genotypic values of the locus with which we are concerned. Thus the distribution of phenotypic values will be the distribution of the sum of, say, normally distributed environmental deviations and perhaps binomially distributed genetic differences. If there is no linkage disequilibrium, epistasis or genotype-environment interaction the distributions of phenotype may only differ in mean. If there are many independent genes of small effect which influence the trait, the phenotypes may be almost exactly normally distributed. The variance will be equal to the total phenotypic variance for the trait, less that actually contributed by the locus under consideration. This approximation has been used in infinite population theory by Griffing (1960) and Latter (1965), and by Kojima (1961) for finite populations.

However, when there are repeated cycles of selection and several loci affect the selected trait it may be difficult to justify the assumption that the transition probabilities of matrix **P**, for example, are stationary. Selection and inbreeding will change the frequencies and variance at each locus, so that the distribution of phenotypic values for a specified genotype, and therefore the selective values, will not remain constant over generations. The extent to which selective values will change in the presence of other loci will, of course, depend on their initial frequencies, effects and linkage relationships. The general tendency would seem to be for selective values to increase as other loci approach fixation as a result of selection or drift. At the same time, as Robertson (1960) has mentioned, the environmental variance may rise due to inbreeding, and may partially compensate for the reduction in genetic variance. Also, it is clear that genes of large effect and low initial frequency of the favourable allele are most likely to be lost from the population in the first few generations. If they survive to this stage their frequency is unlikely still to be low, and they will become fixed eventually. Thus 'decisions' about the fate of such genes, and essentially all genes with relatively large Ns value, are taken in early generations before the phenotypic variance can have changed appreciably, so that a theory developed for single loci may given satisfactory predictions in such cases. It will be less satisfactory for genes of smaller effect when fixation takes longer, but further work on this topic is clearly required.

When selection is practised in finite populations initially in equilibrium tight

linkage leads to an excess of the repulsion phase (Hill & Robertson, 1966; Latter, 1966). Thus a theory based on single loci overestimates the expected selection gain in this case.

SUMMARY

The effect of selection on individual performance for a quantitative trait is studied theoretically for populations of finite size. The trait is assumed to be affected by environmental error and by segregation at a single locus. Exact formulae are derived to predict the change in gene frequency at this locus, initially by finding the probability distribution of the numbers of each genotype selected from a finite population of specified genotypic composition. Assuming that there is random mating and no natural selection the results are extended to describe repeated cycles of artificial selection for a monecious population. The formulae are evaluated numerically for the case of normally distributed environmental errors.

Using numerical examples comparisons are made between the exact values for the predicted change in gene frequency with values obtained using approximate, but simpler, methods. Unless the gene has a large effect (α) on the quantitative trait, relative to the standard deviation of the environmental errors, the agreement between exact and approximate methods is satisfactory for most predictive purposes. The chance of fixation after repeated generations of selection is also evaluated using the exact method, and by means of a diffusion approximation and simple transition probability matrix methods. Except for very small values of population size (N) and large α the results from the diffusion equation agree closely with those from the exact method. Similar results are found from tests made of the prediction from the diffusion equation that the limit is only a function of $N\alpha$ for a given intensity of selection and initial frequency, and that the rate of advance in gene frequency is proportional to $1/N$ for the same set of parameters.

I am grateful to Professors O. Kempthorne and Alan Robertson for their helpful suggestions and comments on the manuscript.

REFERENCES

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proc. Roy. Soc. Edinb.* **46**, 289–305.

Allan, J. S. & Robertson, A. (1964). The effect of initial reverse selection upon total selection response. *Genet. Res.* **5**, 68–79.

Curnow, R. N. & Baker, L. H. (1968). The effect of repeated cycles of selection and regeneration in populations of finite size. *Genet. Res.* **11**, 105–112.

Dempster, E. R. (1955). Genetic models in relation to animal breeding problems. *Biometrika* **11**, 525–536.

Ewens, W. J. (1963). Numerical results and diffusion approximation in a genetic process. *Biometrika* **50**, 241–249.

Griffing, B. (1960). Theoretical consequences of truncation selection based on the individual phenotype. *Aust. J. Biol. Sci.* **13**, 307–343.

Haldane, J. B. S. (1931). A mathematical theory of natural and artificial selection. VII. Selection intensity as a function of mortality rate. *Proc. Camb. Phil. Soc.* **27**, 131–136.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on the limits to artificial selection. *Genet. Res.* **8**, 269–294.

HILL, W. G. & ROBERTSON, A. (1968). The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60** (in Press).

KIMURA, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Statist.* **28**, 882–901.

KIMURA, M. (1958). On the change of population fitness by natural selection. *Heredity* **12**, 145–167.

KIMURA, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

KOJIMA, K. (1961). Effects of dominance and size of population on response to mass selection. *Genet. Res.* **2**, 177–188.

LATTER, B. D. H. (1965). The response to artificial selection due to autosomal genes of large effect. I. Changes in gene frequency at an additive locus. *Aust. J. Biol. Sci.* **18**, 585–598.

LATTER, B. D. H. (1966). The interaction between effective population size and linkage intensity under artificial selection. *Genet. Res.* **7**, 313–323.

PIKE, D. J. (1969). A comparison of two methods for predicting changes in the distribution of gene frequency when selection is applied repeatedly to a finite population. *Genet. Res.* **13**, 117–126.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proc. Roy. Soc.* B **153**, 234–249.

TEICHROEW, D. (1956). Tables of expected values of order statistics and products of order statistics from samples of size 20 and less from the normal distribution. *Ann. Math. Statist.* **27**, 410–426.

3

The rate of selection advance for non-additive loci.

by

William G. Hill

Genet. Res., Camb. (1969), **13**, *pp.* 165–173
With 7 text-figures
Printed in Great Britain

The rate of selection advance for non-additive loci

By W. G. HILL
Institute of Animal Genetics, West Mains Road, Edinburgh, 9

(*Received* 4 *June* 1968)

1. INTRODUCTION

Some selection experiments and breeding programmes are continued for many generations until a limit is reached after which no further progress can be made. The breeder may wish to predict before undertaking the programme how long it will take to reach the limit, or at least get a large part of the way there if the approach is asymptotic. The initial rate of advance can usually be predicted adequately using classical quantitative genetics theory from the heritability and selection differential, but when selection is continued for several generations in finite populations the genetic variance and, consequently, heritability change as the gene frequencies alter with drift and selection. The manner in which these change and thus the rate of advance over many generations of selection depend on the effects, frequency and number of genes influencing the quantitative trait. The breeder and experimentalist may therefore be interested in analysing their results in the hope of obtaining some information about the inheritance of the selected trait.

As a measure of the rate of advance Robertson (1960) defined the half-life of the selection process as the time taken to get half-way to the limit. Robertson (1960) and Hill & Robertson (1966) have given results for the half-lives of single additive genes, and Latter (1966) and Hill & Robertson (1966) have discussed some of the effects of linkage on half-lives with pairs of linked additive loci. The half-life can be measured from practical experiments which are taken to the selection limit, and has the advantage in that it is a statistic whose units are solely generations (or years) and do not involve scale factors such as gallons of milk. Also, if the rate of advance is close to exponential in form the half-life and the total advance completely describe the process. This can occur with additive genes with weak selection (Robertson, 1960) or with dominant genes with no selection, when the decline in the mean is proportional to the inbreeding coefficient.

In this paper the influence of initial gene frequency and size of gene effects will be investigated for genes showing complete dominance. The model used will be limited strictly to single loci, but may be expected to hold approximately for many independent loci with the same initial frequency and effect. Of course this cannot represent the real situation for any quantitative trait, but the results should still be of some diagnostic value. The methods used here for calculating the selection advance in finite populations are only approximate, but they have been shown in

the previous paper by Hill (1969) to give a good fit to exact results calculated for a simple model of population structure.

2. MODEL AND METHOD

Consider a single locus with two alleles, at which the favourable allele A has initial frequency q_0. From the original population a large number of replicated finite subpopulations or lines are drawn, and in each selection is practised. The average frequency of A at generation t is $E(q_0)$ and we let $u(q_0) = \lim_{t\to\infty} E(q_t)$. Thus $u(q_0)$ is the chance of fixation of A, and is the probability that at the selection limit it is fixed in the population. The total gain is $u(q_0) - q_0$ and the half-life is the value of t such that

$$E(q_t) - q_0 = \tfrac{1}{2}[u(q_0) - q_0].$$

Hill (1969) has described the model which we shall approximate. Each generation N out of a total of M monecious individuals are selected on their own value for the quantitative trait. These N individuals mate at random, and random selfing is included. If A is dominant over the alternative allele a the expected change in gene frequency in a line with frequency q is approximately $k\alpha q(1-q)^2$. This formula was originally derived by Kojima (1961); k is the mean of the top N from M order statistics from a standardized normal distribution and α is the difference in genotypic value between AA and aa individuals expressed as a proportion of the standard deviation of phenotypic values about the genotypic value. If A is the only locus affecting the trait α remains constant, and we let $s = k\alpha$. Hill (1969) has shown that a transition probability matrix $\mathbf{B}$ with elements (b_{ij}), $i, j = 0, \ldots, 2N$ given by

$$b_{ij} = \binom{2N}{j}\left[\frac{i}{2N} + s\frac{i}{2N}\left(1-\frac{i}{2N}\right)^2\right]^j\left[1-\frac{i}{2N}-s\frac{i}{2N}\left(1-\frac{i}{2N}\right)^2\right]^{2N-j}$$

is a suitable approximation for the transition probability matrix in which the selection process is described exactly. The element b_{ij} is the (approximate) probability that the N parents contain jA alleles at generation $t+1$ given that they had i at generation t.

Since the matrix $\mathbf{B}$ is easy to compute it has been used for all the results given in this paper. A large value of $N(32)$ has been used, such that k is close in value to i, where

$$i = \lim_{N\to\infty} k,$$

with N/M constant. Thus the selective value of the gene is approximately $i\alpha$ which appears in the well-known formulae for selection in infinitely large populations (e.g. Griffing, 1960). Using a diffusion equation to give a continuous approximation to the selection process it can be shown that the selection limit is a function of only Ns and q_0, and that the mean gene frequency at generation t is a function of the same parameters and also t/N. This simplifying assumption has also been investigated by Hill (1969) and found to be an adequate approximation for most descriptive purposes.

Since we are concerned now with the genetic implication of the results one change in definition will be made compared with the previous paper (Hill, 1969). Selective values will always be assumed to be positive; the allele favoured by selection will have initial frequency q_0 and it will be stated whether this allele is recessive or dominant over its alternative.

In the limiting case as Ns becomes very small an explicit formula for the average gene frequency has been given by Robertson (1960). With the recessive allele favoured

$$E(q_t) = q_0 + Nsq_0(1-q_0)[1-e^{-t/2N} - \tfrac{1}{3}(1-2q_0)(1-e^{-3t/2N})]$$

approximately. This formula was used to compute the half-life of the gene frequency for the limiting value as Ns becomes zero by solving for t with the mean frequency half-way to its limiting value. For larger values of Ns the transition matrix **B** was iterated repeatedly onto a vector of mean gene frequencies. The method is described in detail by Hill (1969).

3. CHANGES IN GENE FREQUENCY

Half-lives for the change in gene frequency are given for a wide range of parameters in Figs. 1–3. Figure 1 shows the case of additive gene action which has been included for comparison and is reproduced from Hill & Robertson (1966). In the additive model the two homozygotes differ in selective value by $s = k\alpha$. The transition matrix used was analogous to the approximate matrix **B** for complete dominance. Rows of the matrix were obtained by a binomial expansion with index $2N$ and mean $(i/2N) + s(i/2N)[1-(i/2N)]$. The allele favoured by selection is a recessive in Fig. 2 and a dominant in Fig. 3.

Further information on the pattern of responses is contained in Table 1. There the ratio of quarter-lives to half-lives $t(\frac{1}{4})/t(\frac{1}{2})$ and the ratio of half-lives to three-quarter lives $t(\frac{1}{2})/t(\frac{3}{4})$ are given for a few values of the parameters Ns, q_0 and mode of gene action. The quarter-life is, of course, the time taken to get a quarter of the way to the limit. If the pattern of advance is exponential, of the form

$$E(q_t) - q_0 = [u(q_0) - q_0](1-e^{-lt})$$

then $t(\frac{1}{4})/t(\frac{1}{2}) = 0{\cdot}415$ and $t(\frac{1}{2})/t(\frac{3}{4}) = 0{\cdot}500$, independent of the value of the constant l. With additivity, as $Ns \to 0$ changes are approximately exponential and $l = 1/2N$ (Robertson, 1960). Small values of $t(\frac{1}{4})/t(\frac{1}{2})$ and $t(\frac{1}{2})/t(\frac{3}{4})$ reflect rapid early advance followed by a prolonged period of relatively slower advance.

With small Ns the half-life of a favourable recessive gene is $2{\cdot}1N$ if q_0 is close to zero, and is $1{\cdot}0N$ if q_0 is close to unity, with intermediate values being obtained for other initial frequencies. When $q_0 = 0{\cdot}5$, the half-life is $1{\cdot}4N$, which is the same as for additive genes with all values of q_0 and small Ns (Robertson, 1960). When the recessive is favoured an increase in Ns always reduces the half-life for any starting frequency, reflecting the fact that for a given value of N the larger Ns then the greater the selective value s and the rate of advance. Similarly, for any specific value of Ns the half-life is always less with higher initial frequency apparently because the favoured allele requires a smaller change in frequency before fixation.

A similar result is obtained for additive genes although the effect is not so pronounced. With recessive genes, when the response is proportional to $sq^2(1-q)$, slow rates of advance are clearly made if the frequency is low.

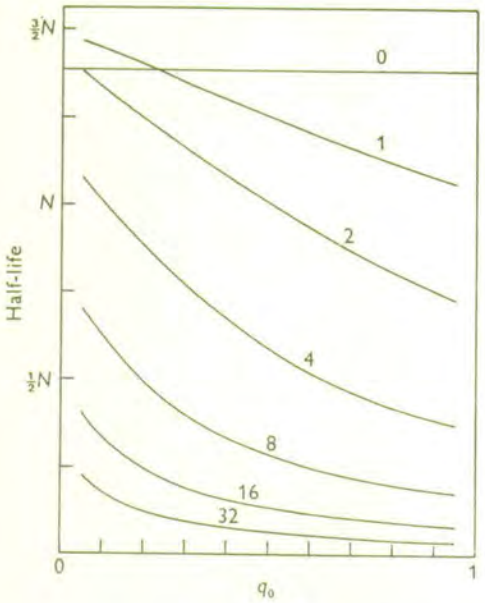

Fig. 1

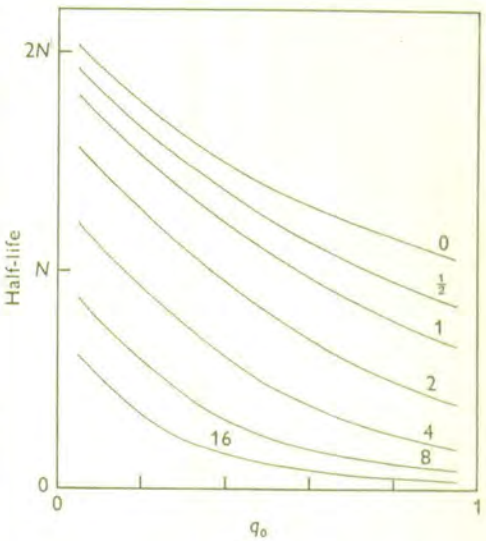

Fig. 2

Fig. 1. Half-life of change in gene frequency with selection for an additive gene. Time is measured in generations, and curves are plotted for different values of Ns with initial frequency q_0. (Reproduced from Hill & Robertson, 1966.)

Fig. 2. As Fig. 1, but with selection for a favourable recessive gene.

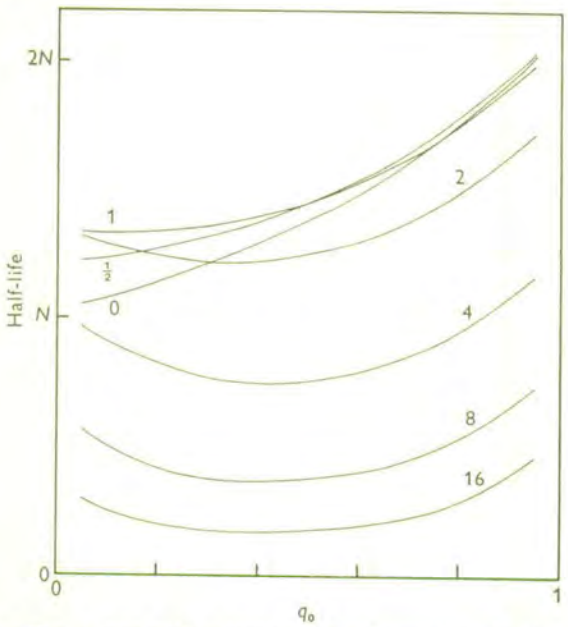

Fig. 3. As Fig. 1, but with selection for a favourable dominant gene.

However, when the dominant allele is favoured we see in Fig. 3 that unless Ns is small the shortest half-lives occur with intermediate initial frequencies, the minima in the curve occurring at about $q_0 = 0{\cdot}4$. It would seem that two counteracting forces cause this minimum. With high initial frequencies the half-life is long, because the rate of advance, proportional to $sq(1-q)^2$ is low at a high frequency of the dominant. On the other hand, with low initial gene frequency the half-life is prolonged because a greater total advance is required before the desired allele is fixed. The latter effect becomes less important with very small Ns values and low q_0 for the favourable allele is rarely fixed, so that there is no minimum in the curve of half-life against q_0 for small Ns when the favourable allele is dominant.

Table 1. *Ratio of quarter-life to half-life* $t(\frac{1}{4})/t(\frac{1}{2})$ *and half-life to three-quarter-life* $t(\frac{1}{2})/t(\frac{3}{4})$ *measured for* $E(q)$ *for various sets of parameters*

	$t(\frac{1}{4})/t(\frac{1}{2})$			$t(\frac{1}{2})/t(\frac{3}{4})$		
Initial frequency (q_0)...	0·1	0·5	0·9	0·1	0·5	0·9
			Additive model			
Ns 0	0·415	0·415	0·415	0·500	0·500	0·500
1	0·442	0·416	0·395	0·528	0·502	0·477
4	0·513	0·436	0·402	0·599	0·521	0·480
16	0·589	0·463	0·419	0·674	0·565	0·506
			Recessive model			
Ns 0	0·512	0·415	0·390	0·571	0·500	0·462
1	0·539	0·428	0·386	0·599	0·510	0·453
4	0 592	0·467	0·406	0·654	0·548	0·487
16	0·659	0·514	0·450	0·697	0·628	0·529
			Dominant model			
Ns 0	0·390	0·415	0·512	0·462	0·500	0·571
1	0·409	0·405	0·495	0·482	0·492	0·555
4	0·471	0·391	0·483	0·529	0·472	0·548
16	0·534	0·386	0·449	0·559	0·424	0·526

When we compare the half-lives for the change in gene frequency for the three models in Figs. 1–3 we find that at intermediate initial frequencies the half-lives are about the same in each case, and for $Ns \to 0$ and $q_0 = 0{\cdot}5$ they become exactly the same. If the desired allele has a low initial frequency the half-life is shortest if it is dominant and longest if it is recessive. By contrast, if the favoured allele has a high initial frequency the ranking is reversed.

4. CHANGES IN THE POPULATION MEAN OF THE QUANTITATIVE TRAIT

We are not usually able to observe changes in gene frequency for dominant genes in single lines, although we can observe changes in the mean of the quantitative trait under selection. However, if crosses are made between replicated lines the response in the mean of the metric trait in crossbred inidividuals will be proportional to the response in gene frequency. With additive genes the pattern of

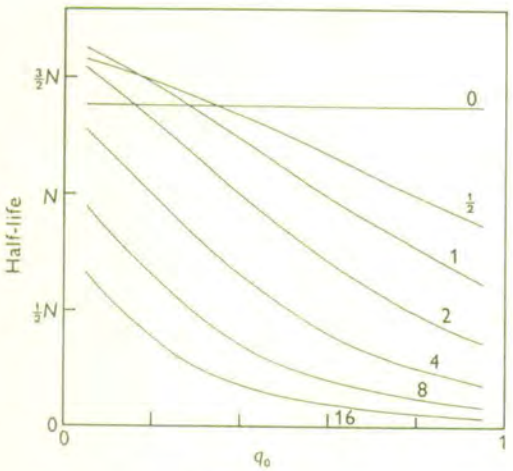

Fig. 4

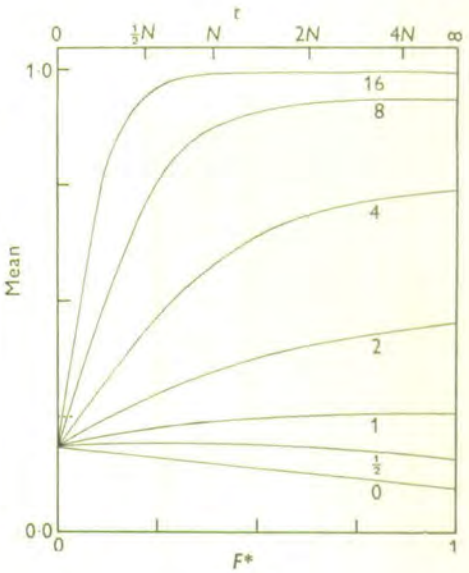

Fig. 5

Fig. 4. Half-life of change in the population mean with selection for a favourable recessive gene. Time is measured in generations and curves are plotted for different values of Ns with initial frequecy q_0.

Fig. 5. The population mean, expressed as the average value of $1-(1-q)^2$, for selection for a favourable dominant gene with initial frequency 0·1. Time is shown in terms of $F^* = 1-\exp(-t/2N)$ (lower scale) and in generations (upper scale). The response is plotted for several values of Ns.

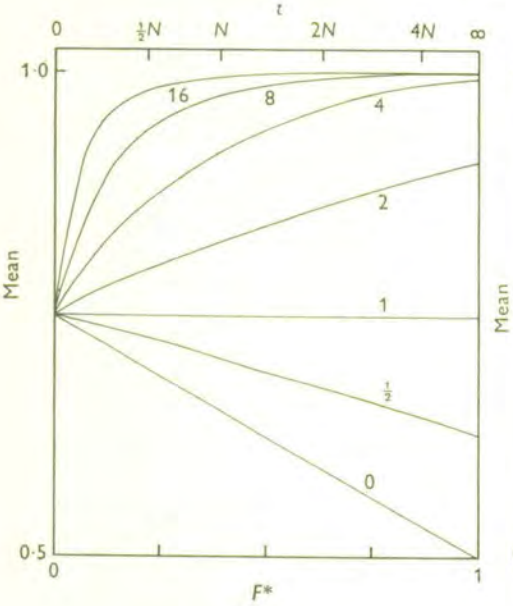

Fig. 6

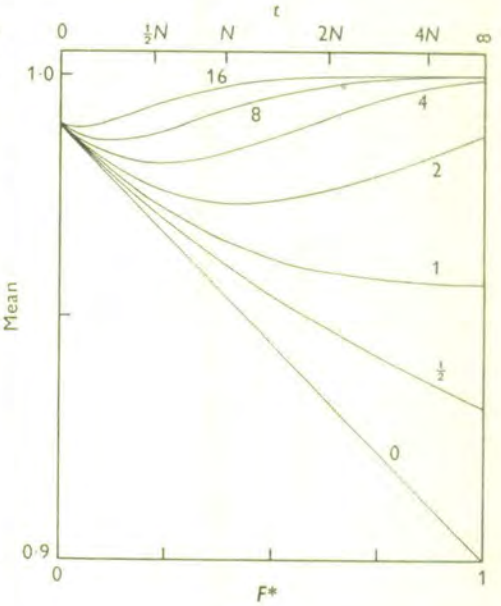

Fig. 7

Fig. 6. As Fig. 5, but with initial frequency 0·5.

Fig. 7. As Fig. 5, but with initial frequency 0·9.

change in the population mean (for the quantitative character) is the same as for the gene frequency, even in a single line, since the mean is a linear function of gene frequency. However the mean of the trait is proportional to q^2 with a recessive, or $[1-(1-q)^2]$ with a dominant gene in a single population. We therefore need to determine the pattern of change in this mean also. Thus the expected value of q_t^2 at each successive generation was computed using the matrix **B** in the same manner as $E(q_t)$ had been computed (Hill, 1969). A half-life for the population mean was also calculated.

The population mean is affected both by selection and by inbreeding depression unless there is additivity, and these effects oppose each other when the dominant allele is favoured. As a result, changes in the mean may be in a direction opposite to that in which selection is practised, or the direction of the response may alter after a few generations of selection, and the half-life is then not a very useful concept. Therefore, while half-lives are shown in Fig. 4 for the population mean of favourable recessives, where inbreeding enhances the advance, some response curves are presented for the case of favourable dominants. In Figs. 5–7 for initial frequencies of 0·1, 0·5 and 0·9 respectively, the expected value of the quantitative trait is plotted against generation number. The scale of generations is transformed to $F^* = 1-e^{-t/2N}$, which is approximately equal to $1-(1-1/2N)^t$, the inbreeding coefficient for neutral genes, and generally provides a suitable contraction of the time scale as t becomes large.

If there is no selection, so that $Ns = 0$, the only force is inbreeding depression and the half-life is reached when $F^* = 0{\cdot}5$, which takes $1{\cdot}4N$ generations for all initial frequencies. At higher Ns values the pattern of half-lives is much the same for both mean and gene frequency when there is selection for the recessive allele. The half-life of the mean for the case of the desirable recessive only exceeds $1{\cdot}4N$ generations when the initial frequency is low and Ns takes intermediate values. Selection then increases the probability of fixation of the favourable allele but requires a long period of selection before fixation because there is a slow initial response with a low-frequency recessive. With higher Ns values fixation of the favourable allele can occur more rapidly.

Selection for a dominant allele is able to counteract the effects of inbreeding sufficiently so that the final mean exceeds the initial mean if Ns is greater than about 0·5 for $q_0 = 0{\cdot}1$, or about 1 for $q_0 = 0{\cdot}5$ and 2 for $q_0 = 0{\cdot}9$. However, at low initial frequencies there may be a period of an advance in the mean, followed by a slight decline, whereas at higher frequencies the decline occurs in the earlier generation. These results reflect the low rate of response obtained with favourable dominants at high initial frequency. In the first few generations there is little additive variance (proportional to $q(1-q)^3$), but this increases as random drift moves the frequency in some replicates to intermediate values (Robertson, 1952; Hill & Robertson, 1968).

5. DISCUSSION

In the model studied we have assumed that only a single locus influences the quantitative trait. However, the results can be expected to hold approximately for several independent loci so long as the selective values of the genes do not change too much as a result of selection. With a heritability of 0·25, for example, the value of the phenotypic standard deviation could decrease by about 10% if all genes went to fixation. Since α is inversely proportional to the phenotypic standard deviation it would correspondingly increase by about 10%. But we see in the graphs that this magnitude of change in α and hence Ns has little qualitative effect on the results. These assumptions have been discussed further by Hill (1969).

The pattern of response we have found is greatly affected by the nature of the gene action. Generally a short half-life, say $N/2$ or less generations, indicates that the gene has a high selective value. Thus if artificial selection is being practised for a quantitative trait which is likely to be affected by several loci, a short half-life may indicate that at least some of the genes have a large effect on the selected trait. However earlier work has shown that the half-lives of additive genes which are initially in linkage equilibrium are usually reduced if they are tightly linked to other genes undergoing selection (Latter, 1966; Hill & Robertson, 1966).

These results have been used in practice to analyse results from selection experiments. Roberts (1966) has calculated the half-life from selection studies on body weight in mice which had been continued till a plateau was reached. From the half-life Roberts estimated the average size of gene effects and the number of loci influencing the selected trait. But with this technique it is necessary to make very strong assumptions about the distribution of initial gene frequencies and effects, usually that they are the same for all loci. If it is also assumed that the loci are independent, when, in fact, some may be closely linked, estimates of gene effects are likely to be biased upwards.

SUMMARY

Expected changes in the gene frequency and the population mean for a quantitative trait are described for selection in a population of size N at a single locus where the favoured allele has initial frequency q_0 and selective value s. Models of additive and completely dominant gene action are compared. Results are generally expressed as the half-life of the total change relative to N.

If the favoured allele is additive or recessive the half-life of the gene frequency and mean of the trait are usually reduced when q_0 or Ns is increased. However, if the dominant allele is favoured the half-life of gene frequency is still generally reduced as Ns is increased, but has a minimum at low or intermediate values of q_0. Since inbreeding depression and selection oppose each other when the dominant allele is favoured the response in the mean of the quantitative trait may change in direction during selection.

REFERENCES

GRIFFING, B. (1960). Theoretical consequences of truncation selection based on the individual phenotype. *Aust. J. biol. Sci.* **13**, 307–343.

HILL, W. G. (1969). On the theory of artificial selection in finite populations. *Genet. Res.* **13**, 143–163.

HILL, W. G. & ROBERTSON, A. (1966). The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294.

HILL, W. G. & ROBERTSON, A. (1968). The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**, (in Press).

KOJIMA, K. (1961). Effects of dominance and size of population on response to mass selection. *Genet. Res.* **2**, 177–188.

LATTER, B. D. H. (1966). The response to artificial selection due to autosomal genes of large effect. III. The effects of linkage on the rate of advance and approach to fixation in finite populations. *Aust. J. biol. Sci.* **19**, 131–146.

ROBERTS, R. C. (1966). The limits to artificial selection for body weight in the mouse. I. The limits attained in earlier experiments. *Genet. Res.* **8**, 347–360.

ROBERTSON, A. (1952). The effect of inbreeding on the variation due to recessive genes. *Genetics* **37**, 189–207.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proc. R. Soc.* B **153**, 234–249.

4

The effects of inbreeding at loci with heterozygote advantage

by

William G. Hill and Alan Robertson

THE EFFECTS OF INBREEDING AT LOCI WITH HETEROZYGOTE ADVANTAGE

W. G. HILL AND ALAN ROBERTSON*

Institute of Animal Genetics, Edinburgh 9

Received May 20, 1968

SINCE the early studies of FISHER and WRIGHT the theory of selection within populations of finite size has received much attention. KIMURA (1964) has reviewed the part of the theory that is based on continuous models in which it is usually assumed that individuals mate at random within small closed sub-populations or lines. In such a situation we are concerned with the distribution of the frequency of individual genes over many replicate lines, or, equivalently, the distribution of the frequency of identical genes within the same line. In this report we study selection favouring heterozygous individuals with random mating within lines and no selection or crossing occurring between lines. The model for inbreeding which we discuss must be distinguished from an alternative situation, perhaps more common in plants, in which inbreeding occurs within an infinitely large population as a result of non-random mating, for example by selfing or mixed selfing and outcrossing. In the latter type of model, selection also may occur between sublines and recurrent mutation is not required for equilibria of gene frequency to occur without fixation, whereas it is in our model. These equilibrium situations have been analysed recently in some detail by ALLARD and co-workers. Many of their results for single loci are reviewed by JAIN and WORKMAN (1967) and analysis of a two locus model is given by JAIN and ALLARD (1966).

The effect of selection for heterozygous individuals in small lines when there is no between-line selection has been studied by REEVE (1955) using transition probability matrices for mating types in lines of only a few individuals, and by ROBERTSON (1962). The latter considered two situations—firstly when there is a balance between mutation and fixation and secondly when, in the absence of mutation, the amount of heterozygosis is declining at a steady rate. In both, the critical factor proved to be the equilibrium gene frequency, which depends on the relative fitness of the two homozygotes. If the equilibrium frequency lies outside the range 0.2 to 0.8 then selection may have an effect opposite to that usually expected and increase the rate of fixation.

In the present paper we shall be concerned with the intermediate stages of selection for the heterozygote in small lines with a known initial gene frequency. Selection may alter the mean gene frequency and the proportion of heterozygotes

* Member of the Agricultural Research Council Unit of Animal Genetics.

as well as the mean of quantitative traits. Two particular situations are of interest. In the context of natural selection, we may be concerned with the effect on fitness due to such loci when a large population is suddenly reduced in size. The initial gene frequency can then be assumed to be that at equilibrium in large populations. We might then speak of the effect of inbreeding on fitness though it is in fact the joint effect of natural selection and inbreeding.

The second situation is one involving artificial selection in which the heterozygote has an advantage because it is the best of the three genotypes for the character under selection. We now have no reason to assume any particular initial gene frequency.

MODEL

We shall consider only the case of two alleles at a single locus at which there is no mutation and which is not linked to other loci under selection. Let us assume that the relative selective advantages of the genotypes A_1A_1, A_1A_2 and A_2A_2 are $1-s_1$, 1 and $1-s_2$, respectively. Let q denote the frequency of the A_1 allele and let $\bar{q}$, given by $\bar{q} = s_2/(s_1+s_2)$, be the equilibrium frequency in large populations. We shall only consider values of $\bar{q} \leqslant 0.5$ since there is symmetry about this point.

Transition matrix for monecious individuals: In the model which we shall investigate in most detail we assume that there are non-overlapping generations and that the parents comprise N monecious individuals which undergo random mating including random selfing. At some generation t let there be i A_1 alleles among the $2N$ alleles at the A locus in the adults, where $0 \leqslant i \leqslant 2N$. For brevity let $q = i/2N$. The genotypic frequencies among the zygotes at generation $t+1$ will be q^2, $2q(1-q)$ and $(1-q)^2$ for A_1A_1, A_1A_2 and A_2A_2 individuals, respectively. These genotypic frequencies depend only on the gene frequencies in their parents since there is random mating. Assuming that selection acts through differences in viability, the N individuals which become parents of the next generation will have a multinomial distribution of genotypic frequencies.

$$f_i(x,y,z) = \binom{N}{x\ y\ z} \left\{\frac{q^2(1-s_1)}{\bar{w}}\right\}^x \left\{\frac{2q(1-q)}{\bar{w}}\right\}^y \left\{\frac{(1-q)^2(1-s_2)}{\bar{w}}\right\}^z$$

where $f_i(x,y,z)$ is the probability that there are x A_1A_1, y A_1A_2 and z A_2A_2 individuals surviving, with $q = i/2N$. The average fitness, $\bar{w}$, is

$$\begin{aligned}\bar{w} &= 1 - s_1q^2 - s_2(1-q)^2 \\ &= 1 - (s_1+s_2)\ [\bar{q}(1-\bar{q}) + (q-\bar{q})^2] \qquad (1)\end{aligned}$$

The probability that the N survivors have exactly j A_1 alleles is given by summation of the probabilities of all combinations of genotypic frequencies for which $2x+y=j$. Thus we obtain p_{ij}, the probability that there are j A_1 alleles at generation $t+1$ given that there were i at generation t as

$$p_{ij} = \sum_{\substack{2x+y \\ =j}} f_i(x,y,z),\ i,j = 0, \ldots, 2N$$

We let $\mathbf{P}$ be the transition probability matrix with elements p_{ij}. Since the selective

values are assumed to be independent of generation, **P** is independent of generation number, t, also.

The expected change in gene frequency, $\delta q = E(j/2N - q | q = i/2N)$, is

$$\delta q = -(s_1 + s_2) q(1-q)(q - \bar{q})/\bar{w}$$

which is the usual formula for response with a model of heterozygote advantage.

The expected fitness, gene frequency and heterozygosity were computed in successive generations by repeated multiplication using the transition matrix. For example, let the expected gene frequency at generation t, conditional on the initial frequency being q_0, be $E(q_t|q_0)$ or simply $E(q)$. Also let $\mathbf{v}_{(t)}$ be a vector with elements $v_{i(t)} = E(q_t | q_0 = i/2N)$.

Thus

$$v_{i(0)} = i/2N, i = 0, \ldots, 2N,$$

and we obtain

$$\mathbf{v}_{(1)} = \mathbf{P}\mathbf{v}_{(0)},$$
$$\mathbf{v}_{(2)} = \mathbf{P}\mathbf{v}_{(1)}$$

and, in general,

$$\mathbf{v}_{(t)} = \mathbf{P}\mathbf{v}_{(t-1)} \tag{2}$$

Iteration of (2) was repeated on a computer for up to $t = 8N$ generations, but was terminated earlier if there was almost complete fixation or the distribution of gene frequency among lines still segregating appeared to reach a state of steady decline. Then the value of λ, given by

$$\lambda = [v_{i(t)} - v_{i(t-1)}]/[v_{i(t-1)} - v_{i(t-2)}]$$

is constant for sufficiently large t and all i, $0 \leqslant i \leqslant 2N$. Results for later generations were obtained by assuming that the steady state had been reached, computing the dominant non-unit latent root, λ, and using this to predict subsequent changes. The expected heterozygosity, $E[2q(1-q)]$ within lines was computed in a similar manner, with $v_{(i)0}$ becoming $2(i/2\mathrm{N})(1 - i/2N)$. The mean fitness (equation 1) is a linear function of $(q - \bar{q})^2$, with high values denoting low fitness, and, for simplicity, fitness has been expressed in this form. Since

$$E[(q - \bar{q})^2] = \bar{q}^2 + (1 - 2\bar{q})E(q) - E[q(1-q)] \tag{3}$$

the expected fitness could be evaluated from the expected gene frequency and heterozygosity.

Transition matrix for diecious individuals: Although the model with monecious individuals lends itself to simple numerical evaluation on a computer, it does not represent the real situation in most species. A model with two distinct sexes and random mating between the two sexes was therefore investigated for small values of population size, with other model assumptions as in the monecious case. In general there are $(2N_m + 1)(2N_f + 1)$ possible states of gene frequency in the two sexes if there are N_m males and N_f females in each replicate line every generation. We shall only consider the case where the population sizes and selective values are the same for the two sexes, and we let $N_m = N_f = L$ and a state i specify that there are i_m A_1 alleles in male parents and i_f A_1 alleles in female parents, $(0 \leqslant i_m, i_f \leqslant 2L)$. With random mating, the zygotic frequencies in the progeny from parents in state i are

$$q_m q_f \, A_1 A_1, \; q_m(1-q_f) + (1-q_m)q_f \, A_1A_2, \; (1-q_m)(1-q_f) \, A_2A_2$$

where $q_m = i_m/2L$, $q_f = i_f/2L$. The mean gene frequency among the progeny is $q = (q_m + q_f)/2$, and letting $r = (q_m - q_f)/2$, the proportion of heterozygotes becomes $2q(1-q) + 2r^2$. Of course, the states could be defined in terms of q and r among the zygotes, instead of q_m and q_f in their parents. The mean fitness in terms of q and r is

$$\overline{W} = 1 - (s_1 + s_2)\,[\bar{q}(1-\bar{q}) + (q-\bar{q})^2 - r^2] \tag{4}$$

With selection operating independently in males and females, the probability b_{ij} that a line is in state j at generation $t+1$, given that it was in state i at generation t is

$$b_{ij} = P(j_m \mid i).\; P(j_f \mid i)$$

where $P(j_m \mid i)$, $P(j_f \mid i)$ are the marginal probabilities of obtaining j_m A_1 alleles in males and j_f A_1 alleles in females, respectively. For example,

$$P(j_m \mid i) = \sum_{\substack{2x+y \\ = j_m}} \binom{L}{x\,y\,z} [q_m q_f(1-s_1)]^x [q_m(1-q_f) + (1-q_m)q_f]^y \cdot [(1-q_m)(1-q_f)(1-s_2)]^z / \overline{W}^L \tag{5}$$

and, since the selection coefficients are assumed to be the same in males and females, $P(j_f \mid i)$ is obtained by substituting j_f for j_m in (5). The expected change in gene frequency, δq, is in one generation

$$\delta q = -\frac{(s_1+s_2)q(1-q)(q-\bar{q}) + r^2(s_1+s_2)(q+\bar{q}-1)}{\overline{W}} \tag{6}$$

If terms of order $(r^2)^a(s_1+s_2)^b$ are ignored if $a+b>2$, $\overline{W}$ in equation (4) may be replaced by the $\bar{w}$ of equation (1) relating to the monecious model.

Some simplification of the matrix **B** with elements b_{ij} is possible because of symmetry. It is shown in the APPENDIX that iteration can be performed with a vector of dimension $(2L+1)(L+1)$ and a square matrix of the same dimension, rather than with a vector and matrix of dimension $(2L+1)^2$. Even so, it was necessary to restrict computation to matrices with $L = 5$ giving an effective population size of 10, whereas with the model with only one sex it was possible to work with N as large as 40.

Continuous model approximation: As N becomes infinitely large, but with $N(s_1+s_2)$ remaining finite, the selection process can be approximated by a continuous model using a diffusion equation (WATTERSON 1962; KIMURA 1964). The KOLMOGOROV forward equation has not been solved explicity, although the dominant latent root has been evaluated (MILLER 1962). However, we can use the form of the equation to make generalisations about our results, since the inbreeding and selection process becomes only a function of $N(s_1+s_2)$, $\bar{q}$ and the initial frequency, q_0, so long as time is measured on a scale proportional to N. Tests were made to find the adequacy of this generalisation for small values of N, and results are shown in Table 1. The linear function of fitness, $E[(q-\bar{q})^2]$, is tabulated for various values of $N(s_1+s_2)$, $\bar{q}$ and t/N generations, for $N = 10$, 20 and 40 in the monecious model. The results obtained with different values of N are seen to be very similar, except with the largest value of s_1+s_2 (0.8). Also,

TABLE 1

$E[(q-\bar{q})^2] \times 10^5$ *computed for several values of population size* (N) *with a monecious model (M), and for 5 males and 5 females* (N = 10) *with a diecious model (D).*

$q_0 = \bar{q}$, *except for* $\bar{q} = 0.0$ *when* $q_0 = 0.1$.

$\bar{q}$	Generations $\times 1/N$	$N(s_1+s_2)$ Method (N)	1				4				16	
			D(10)	M(10)	M(20)	M(40)	D(10)	M(10)	M(20)	M(40)	M(20)	M(40)
0.0	0.5		2654	2665	2638	2625	1924	1809	1803	1800	558	581
	1.0		3488	3472	3450	3440	1687	1495	1541	1563	205	220
	2.0		4068	4001	4019	4025	892	716	789	828	27	30
	4.0		4195	4115	4172	4199	208	141	179	203	0	1
	$\rightarrow\infty$		4166	4095	4178	4218	9	9	25	37	0	0
0.1	0.5		1841	1883	1857	1845	1528	1498	1487	1481	841	865
	1.0		2969	3009	2982	2969	1914	1923	1856	1871	929	950
	2.0		4083	4093	4090	4088	1733	1615	1697	1739	977	984
	4.0		4844	4820	4852	4866	1288	1219	1292	1335	998	999
	$\rightarrow\infty$		5312	5235	5293	5320	1046	1046	1095	1127	1000	1000
0.3	0.5		4213	4372	4336	4319	3203	3326	3399	3429	1348	1560
	1.0		7148	7418	7378	7357	4711	4919	5105	5188	1708	2006
	2.0		11071	11424	11398	11384	6355	6659	6997	7155	2344	2781
	4.0		15163	15465	15470	15471	8083	8434	8903	9131	3451	4083
	$\rightarrow\infty$		18530	18470	18518	18542	10545	10571	11127	11403	9000	9000
0.5	0.5		4980	5193	5160	5143	3626	3831	3964	4020	1162	1461
	1.0		8588	8982	8943	8922	5471	5875	6182	6316	1179	1522
	2.0		13828	14401	14356	14333	8171	8895	9455	9701	1184	1552
	4.0		19796	20339	20301	20282	12366	13463	14272	14622	1195	1602
	$\rightarrow\infty$		25000	25000	25000	25000	25000	25000	25000	25000	25000	25000

a doubling of N from 20 to 40 has less effect than a doubling from 10 to 20, and, of course, the results must converge to the diffusion equation result as N becomes infinite. Thus, in order to describe the situation, it seems satisfactory to use results from just one value of N. Since we are in practice likely to be interested in values of N much larger than we can handle on the computer, we have only analyzed results obtained with $N = 40$, our largest value. Satisfactory agreement between diffusion equation and exact methods has been found in earlier studies by EWENS (1963), who also derived correction terms for approximations (EWENS 1964). However, these results were for a haploid model with additive selective advantages.

Also included in Table 1 are some results obtained with the diecious model, using the same parameters but with $L = 5$, equivalent to $N = 10$. The function plotted is $E(q_m + q_f)/2 - \bar{q}]^2$ for comparison with the monecious model, but the mean fitness in the diecious model is also affected by departures from Hardy-Weinberg equilibrium due to gene frequency differences between the sexes of the parents. Initially it is assumed that there is the same gene frequency in each sex, so that, for example, with $q_0 = 0.1$, $i_m = i_f = 1$. There is generally adequate

agreement between the one and two-sex models with $N = 10$. At larger values of N better correspondence can be anticipated, since the main effect of having different sexes would appear to be that this causes departure from Hardy-Weinberg equilibrium among the progeny when the parents have different frequencies. However, $E[(q_m - q_f)^2] = 4E(r^2)$ is inversely proportional to N, and so, for large N will become relatively unimportant in the prediction of change of gene frequency (equation 5). FELDMAN (1966) uses the theory of WATTERSON (1962) to show that the same diffusion equation approximates both models. We therefore seem justified in drawing conclusions about populations with random mating between two sexes from our results with populations in which there is only one sex with random mating, including random selfing.

SELECTION FROM INITIAL GENE FREQUENCY EQUILIBRIUM

When a large population is suddenly reduced in size, the initial frequency at those loci in which segregation had been maintained in the population by superior fitness of the heterozygote may be assumed to be close to the equilibrium frequency, $\bar{q}$. This situation is clearly of importance and we shall consider it first and in some detail.

The effect on mean fitness, which is the character under selection, can be calculated from the average value of $(q - \bar{q})^2$. It proves useful to use as a modified time scale, $1 - e^{-t/2N} = F^*$, which is approximately equal to the inbreeding coefficient measured from pedigrees. In the absence of selection, the heterozygosity declines as $1 - F^*$. The mean of the character under selection is plotted in Figure 1 for a range of $N(s_1 + s_2)$ values for $\bar{q} = 0.1$, 0.3 and 0.5. For comparison the curve for the expected value of a character controlled by recessives ($\bar{q} = 0$) is also included, with initial recessive frequency 0.05. Figures 2 and 3 show the average heterozygosity and the average gene frequency, respectively, during the inbreeding and selection process. Results for $\bar{q} = 0.5$ are not included in Figures 2 and 3 since for $\bar{q} = q_0 = 0.5$, there can be no change in the mean gene frequency when starting from equilibrium, and thus $E(q) = 0.5$ and $E[2q(1 - q)] = 0.5 - 2E[(q - \bar{q})^2]$ for all t. Some loci may show heterozygote superiority for fitness and yet have additive effects on some observed metric trait. Changes in the mean of this character will therefore be a linear function of the mean gene frequency, $E(q)$, shown in Figure 3.

When the inbreeding is so rapid that selection has very little effect (as might happen for instance by using special crossing programmes in Drosophila) the mean fitness declines linearly with F^* in all cases ($N(s_1 + s_2) = 0$). Some aspects of the results are rather surprising, when it is borne in mind that heterozygote superiority has its greatest effect in maintaining segregation in small populations when the equilibrium frequency is 0.5. However it can be seen that at low values of F^* and $N(s_1 + s_2)$, selection has a greater effect when the equilibrium frequency is 0.1. Figures 2 and 3 show that we are here dealing with two quite separate phenomena which may act in opposite directions in particular cases.

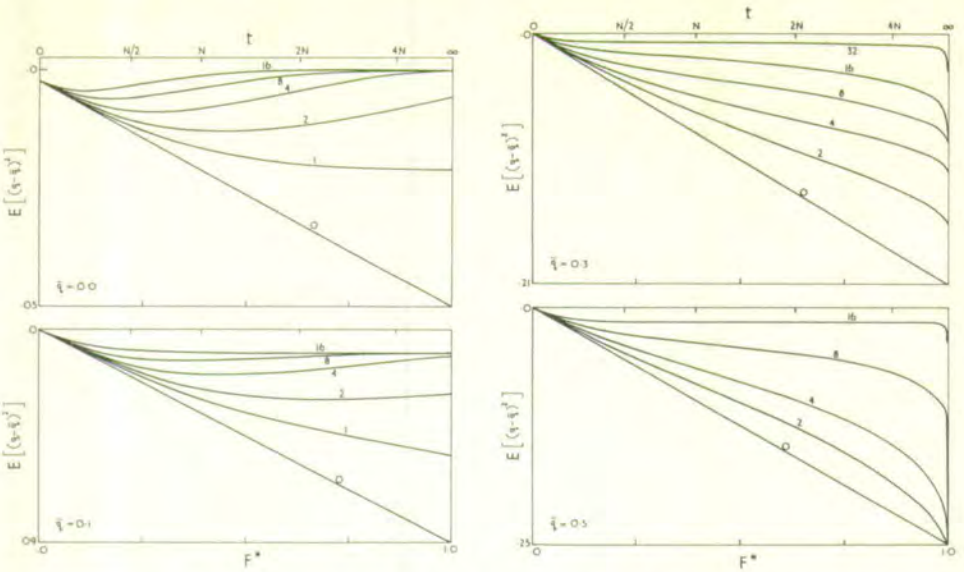

FIGURE 1.—The effect on the mean value of the selected trait $E[(q-\bar{q})^2]$ is plotted against $F^* = 1-e^{-t/2N}$ for values of $\bar{q}$ of 0.0, 0.1, 0.3 and 0.5. The initial frequency equals $\bar{q}$, except for $\bar{q} = 0.0$ (recessive) when the initial frequency is 0.05. Curves are plotted for several values of $N(s_1 + s_2)$.

These correspond to the last two terms in equation (3) which represent respectively the change in gene frequency and in heterozygosity. When the equilibrium gene frequency equals 0.5, the mean gene frequency does not change, so that changes in the mean of the selected character and in the level of heterozygosity are proportional to each other. Inbreeding decline is reduced by the maintenance of a high level of heterozygosity since the second term in (1) vanishes. On the other hand, when $\bar{q} = 0.1$, we find that selection now *reduces* the heterozygosity but at the same time reduces the effect on the mean of the character by preventing fixation of the poorer homozygote. Thus the effect on the mean is due to completely different phenomena in the two cases.

Considering separately the curves for the different equilibrium gene frequencies as $\bar{q}$ approaches zero (including the recessive case as the most extreme value), the inbreeding decline may be halted after a certain time and the selected character then rises again. When there is heterozygote superiority and the initial gene frequency equals $\bar{q}$, the final mean can never be as high as that at the outset since at complete fixation ($F^* = 1$) we cannot do better than fix all populations for the better homozygote, which is inferior to the mean in the initial population at equilibrium. On the other hand, with a single recessive gene in which segregation is maintained by mutation, the final mean may be above the initial value, due to complete exclusion of the recessive at larger values of Ns_1. It is known that

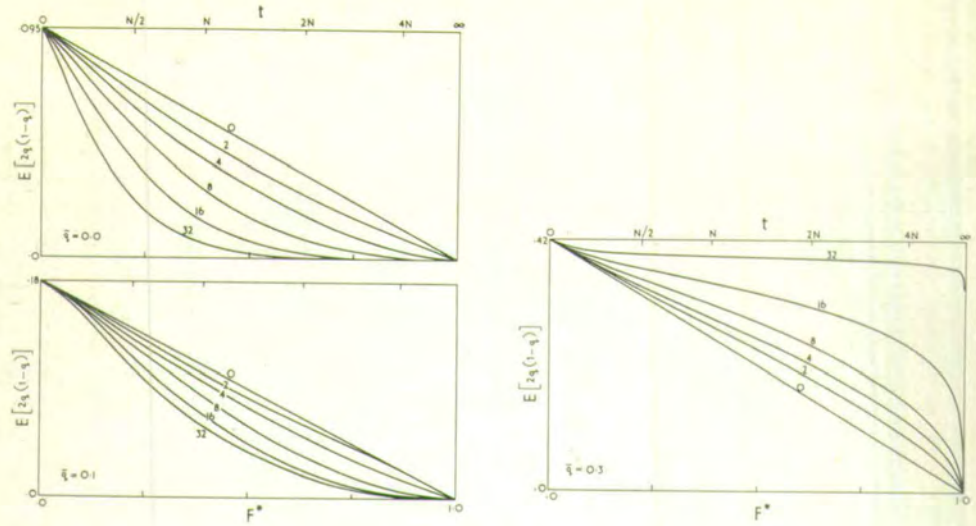

FIGURE 2.—The mean heterozygosity, $E[2q(1-q)]$, plotted as for Figure 1, with $\bar{q} = 0.5$ excluded.

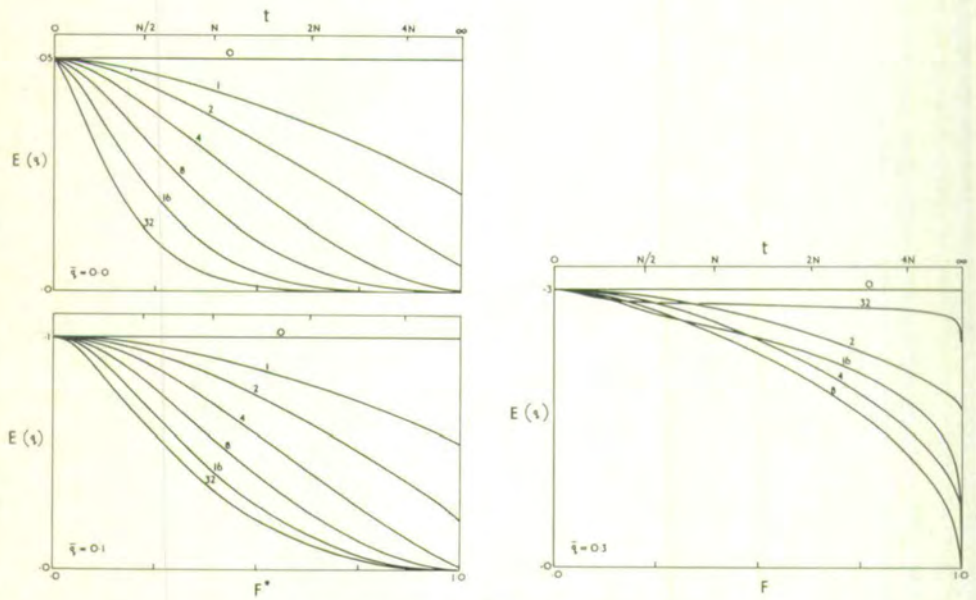

FIGURE 3.—The mean gene frequency, $E(q)$, plotted as for Figure 1, with $\bar{q} = 0.5$ excluded.

in the absence of selection the additive genetic variance within populations due to an initially rare recessive gene will increase up to inbreeding coefficients of 0.5 (ROBERTSON 1952). This provides an explanation of the curves for the extreme values of $\bar{q}$. The first effect of the reduction in population size is a decline in the mean due to an increase in the proportion of homozygotes and selection does not become effective until the additive genetic variance has increased as a result of the spread of gene frequencies. The inbreeding decline is then halted and the mean increased.

When the equilibrium gene frequency is 0.5 and $N(s_1 + s_2)$ is high the mean of the character under selection becomes almost constant after a few generations. There is thereafter a very slow approach to fixation. Selection for the heterozygote under these conditions retards rather than prevents fixation and ultimately all replicate lines will become fixed. As F^* approaches unity the time scale in generations is very much contracted and very small changes in F^* lead to relatively large changes in $E[(q - \bar{q})^2]$.

The curves for $\bar{q} = 0.3$ have a pattern in between those for the other values. In the earlier generations they are similar to those for $\bar{q} = 0.5$, but the effect of selection can be seen from the gene frequency at fixation which is much higher when $N(s_1 + s_2)$ is large and only the better homozygote is fixed. The mean gene frequency (Figure 3) for $\bar{q} = 0.3$ declines more rapidly for $N(s_1 + s_2) = 8$ than for $N(s_1 + s_2) = 32$ since fixation occurs earlier, but the limiting value of $E(q)$ is almost the same in each case.

INITIAL GENE FREQUENCY NOT AT THE EQUILIBRIUM VALUE

There are several situations in which the initial gene frequency may not be at equilibrium. In natural populations there may be a change in environment, which alters the relative fitness of the genotypes, coinciding with a reduction in population size, or there may be departures from equilibrium resulting from random drift at previous reductions in population size. In populations of plants and domestic animals, artificial selection may be applied to a trait which had not previously been important. The selective values (s_1, s_2) are then approximately equal to linear functions of the average genotypic values for the quantitative trait. The adequacy of this approximation for study of truncation selection in infinite populations has been investigated by HILL (1969) and found suitable for most descriptive purposes. We shall illustrate the effects of departures from initial equilibrium for only one value of $N(s_1 + s_2)$, from which we can readily infer the results in other situations.

In Figures 4 and 5 the mean of the quantitative trait, expressed as $[(q - \bar{q})^2]$, and the average gene frequency, respectively, are plotted as a function of initial frequency and generations of inbreeding for $N(s_1 + s_2) = 8$ and $\bar{q} = 0.3$ and 0.5. The curves are drawn for values of t such as 0, $N/2$, N, 2N, $8N$ and ∞ generations, corresponding to $F^* = 0$, 0.22, 0.39, 0.63, 0.98 and 1, respectively.

When $\bar{q} = 0.5$ and $q_0 < 0.5$, there is an initial period of advance in the mean

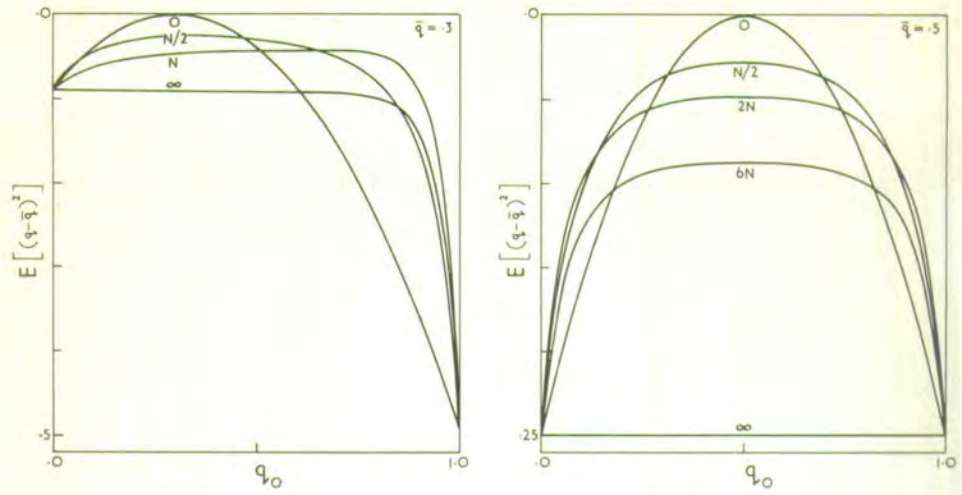

FIGURE 4.—$E[(q-\bar{q})^2]$ for the selected trait is shown as a function of initial frequency for $\bar{q} = 0.3$ and 0.5 and $N(s_1 + s_2) = 8$. Curves are plotted for different numbers of generations.

of the selected trait (Figure 4). This is later lost, and the final mean is the same for all values of q_0 since both homozygous genotypes have the same value. By contrast, when the equilibrium frequency differs from 0.5, the final mean depends on the relative proportions of the two homozygotes fixed, and therefore on q_0. When $\bar{q}$ is less than 0.5 and q_0 high, most of the early advance is retained and there is an overall advance in the mean if $q_0 > 2\bar{q}$, approximately.

If the equilibrium frequency is 0.5, selection always changes the average gene frequency (Figure 5) towards 0.5 until the steady state is reached. After this the average gene frequency remains constant, because the distribution of unfixed

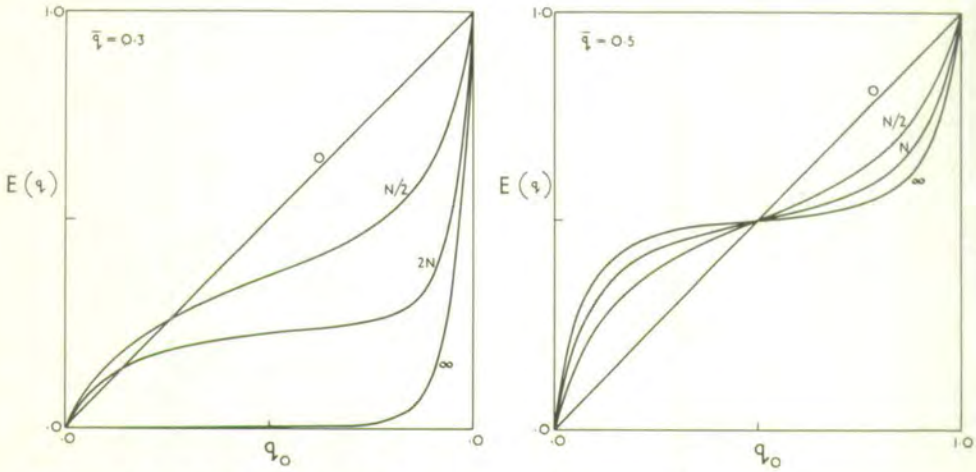

FIGURE 5.—As Figure 4, but a plot of the mean gene frequency, $E(q)$.

classes is symmetric and fixation takes place at the same rate for each homozygote. The mean of the selected trait must then slowly decline. However, if $\bar{q} < 0.5$ and $q_0 > \bar{q}$, selection reduces the gene frequency throughout the inbreeding and selection process. If $q_0 < \bar{q} < 0.5$, there is an initial increase in the average gene frequency as selection towards the equilibrium frequency occurs. The gene frequency distribution of unfixed classes is asymmetric and the poorer homozygote is rarely fixed so that the average gene frequency then declines. We have the interesting phenomenon of unidirectional selection in which there is a reversal of the direction of gene frequency change during selection.

The average gene frequency within segregating lines: Only very large populations can remain at the equilibrium frequency for long periods of time. If we wish to estimate the equilibrium frequency at a locus, as an indirect way of measuring the relative fitnesses of the homozygotes, we may have to use information from populations of finite size. The observed frequency within such populations might be thought to be a good estimator of the equilibrium frequency providing there has been no recent change in environment or immigration. However, the following discussion will show that the observed frequency in small populations is a biassed estimator of the equilibrium frequency in large populations.

Consider an infinitely large population in equilibrium for a locus with heterozygote advantage, from which many identical sub-lines are drawn. The average gene frequency for all lines can be predicted from Figure 3, but this combines data from two types of populations: those which are already fixed, in which the gene frequency is 0 or 1, and those still segregating. We are concerned here with the average frequency within these *segregating* populations, which will reach a steady state value denoted by $\tilde{q}$.

In Figure 6 the relation between $\tilde{q}$ and $\bar{q}$ is plotted for a range of $N(s_1 + s_2)$ values, where results were obtained using the transition probability matrix method described earlier. When $N(s_1 + s_2)$ is infinitely large, the rate of fixation will be very low and $\tilde{q}$ will equal $\bar{q}$. On the other hand, when $N(s_1 + s_2)$ approaches zero, the unfixed classes have a uniform distribution and $\tilde{q} = 0.5$. When $\bar{q} = 0.5$, the distribution of unfixed classes is symmetric about 0.5 for all values of $N(s_1 + s_2)$ so that $\tilde{q} = \bar{q} = 0.5$. When $\bar{q} \neq 0.5$, we know that at final fixation the gene frequency is changed towards that of the better homozygote. If $\bar{q} = 0.3$, for instance, the first effect of the small population size is to spread the gene frequencies about this value. But only those lines with very low frequencies are likely to be fixed so that those left segregating will have a mean gene frequency greater than 0.3. We see from the figure that $\tilde{q}$ almost always lies between $\bar{q}$ and 0.5. With intermediate $\bar{q}(\neq 0.5)$ and $N(s_1 + s_2)$ very large ($\geqslant 16$), $\tilde{q}$ may in some cases lie just outside this range. Here fixation occurs very slowly and the effect can be attributed to the asymmetry of the effect of selection. If the gene frequency drifts from $\bar{q} = 0.3$, say, it is selected more rapidly back when the drift is towards one-half than when it is towards zero because of the term $q(1 - q)$ in

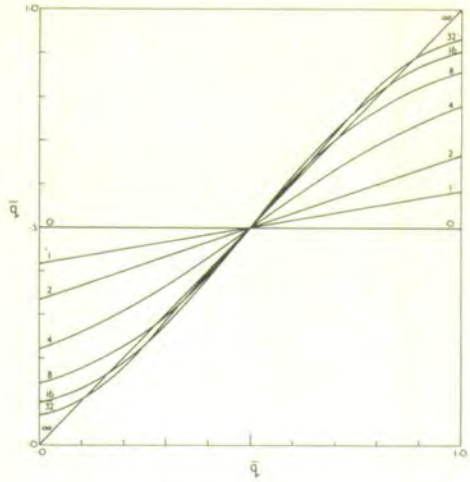

FIGURE 6.—The relations between the average frequency within segregating lines at the steady state, $\bar{q}$, and the equilibrium frequency for large populations, $\bar{q}$. Curves are plotted for several values of $N(s_1 + s_2)$.

$\delta q = -(s_1 + s_2)\ q(1-q)\,(\bar{q} - \bar{q})/\bar{w}$ so that there is a relative excess of populations with extreme gene frequencies.

For most combinations of effective population size and selective values, the average gene frequency within segregating lines is seen to be biassed towards 0.5. Thus if we search for polymorphism within a single small population, we are not likely to find gene frequencies at extreme values. We are then not entitled to infer the relative selective advantages at the loci we observe.

DISCUSSION

A wide variety of consequences of inbreeding is possible when there are loci with heterozygote advantage. Perhaps the most interesting result is that inbreeding depression may be delayed for quite different reasons, depending on the equilibrium frequency, when the population is initially at equilibrium. If the population is not initially at equilibrium, the mean of the selected trait may rise initially and then fall as inbreeding progresses, as well as the reverse.

Two processes were found to reduce inbreeding decline from loci with heterozygote advantage. When the equilibrium frequency was near 0.5, this was due to the maintenance of heterozygosity whereas at extreme equilibrium frequencies it was caused by preferential fixation of the better homozygote. It might be possible to differentiate between these situations in two ways. In the first, lines which had been inbred slowly up to, say, $F^* = 0.75$ (calculated from pedigrees) could then be inbred very rapidly, perhaps by full sibbing. With an equilibrium frequency of one-half, a rapid decline in fitness would be expected to accompany fixation. However, with extreme equilibrium values, most loci will already be

fixed, so little further decline in the mean would be expected. The second method of differentiation has been mentioned in a different context by ROBERTSON (1962), and would apply to very highly inbred replicate lines from the same initial population. Crosses between these lines should show heterosis for loci with intermediate equilibrium frequencies, since both types of homozygote will be fixed in different lines, but for extreme values of the equilibrium frequency most lines will be fixed for the same allele, and no heterosis will be found.

SUMMARY

A theoretical study has been made of the process of inbreeding at loci with heterozygote superiority. Results were obtained using transition probability matrices for monecious and diecious random mating sub-populations, and these alternative models were compared numerically. It was found that, by a suitable choice of parameters, general conclusions drawn from one population size with a monecious model could be applied to other values of population size and to the diecious model.—The rate of inbreeding depression at these loci can be much reduced by selection, but selection is found to act in different ways, depending on the equilibrium frequency in large populations. If this is close to one-half, the effect is due to the maintenance of heterozygosity. With extreme values of the equilibrium frequency it is due to increased fixation of the better homozygote, and this may cause an increase in the mean after a depression during the initial generations of inbreeding.—The relationship between average gene frequency within segregating populations at the steady state and the equilibrium frequency is investigated. This average frequency usually lies between the equilibrium frequency and one-half, giving the impression of more nearly equal selective values for the two alternative homozygotes than is really the case.

LITERATURE CITED

EWENS, W. J., 1963 Numerical results and diffusion approximations in a genetic process. Biometrika **50**: 241–249. —— 1964 The pseudo-transient distribution and its uses in genetics. J. Appl. Prob. **1**: 141–156.

FELDMAN, M. W., 1966 On the offspring number distribution in a genetic population. J. Appl. Prob. **3**: 129–141.

HILL, W. G., 1969 On the theory of artificial selection in finite populations. Genet. Res. (in press).

JAIN, S. K., and R. W. ALLARD, 1966 The effects of linkage, epistasis, and inbreeding on population changes under selection. Genetics **53**: 633–659.

JAIN, S. K., and P. L. WORKMAN, 1967 Generalised F-statistics and the theory of inbreeding and selection. Nature **214**: 674–678.

KIMURA, M., 1964 Diffusion models in population genetics. J. Appl. Prob. **1**: 177–232.

MILLER, G. F., 1962 The evaluation of eigenvalues of a differential equation arising in a problem in genetics. Proc. Camb. Phil. Soc. **58**: 588–593.

REEVE, E. C. R., 1955 Inbreeding with homozygotes at a disadvantage. Ann. Human Genet. **21**: 277–288.

ROBERTSON, A., 1952 The effect of inbreeding on the variation due to recessive genes. Genetics **37**: 189–207. —— 1962 Selection for heterozygotes in small populations. Genetics **47**: 1291–1300.

WATTERSON, G. A., 1962 Some theoretical aspects of diffusion theory in population genetics. Ann. Math. Statist. **33**: 939–957.

APPENDIX

Reduction of the transition matrix **B** *for the diecious model*

Order the $(2L+1)^2$ states of **B** into 3 groups:

Group (1): $2L+1$ states with $i_m = i_f$

Group (2): $\binom{2L+1}{2} = 2L^2 + L$ states with $i_m < i_f$ ordered, for example, as

$$(i_m, i_f) = (0,1), (0,2), \ldots, (2L-1, 2L)$$

Group (3): $2L^2 + L$ states with $i_m > i_f$ ordered similarly to group (2) as

$$(i_m, i_f) = (1,0), (2,0), \ldots, (2L, 2L-1).$$

Since q_m and q_f and thus i_m and i_f can be interchanged in equation (5) and (6), **B** may be partitioned as follows:

$$\mathbf{B} = \begin{pmatrix} \mathbf{C} & \mathbf{D} & \mathbf{D} \\ \mathbf{E} & \mathbf{G} & \mathbf{G} \\ \mathbf{E} & \mathbf{G} & \mathbf{G} \end{pmatrix}$$

where, for example, **C** specifies transitions from states in group (1) to other states in group (1) and is square of dimensions $2L+1$. In order to compute expectations of functions such as the mean gene frequency, $v_{i(t)} = E[(i_m + i_f)/2L \mid \text{initial state} = \text{i}]$, which are symmetric in i_m and i_f, we partition the vector

$$\mathbf{v}'_{(t)} = (\mathbf{x}'_{(t)}, \mathbf{y}'_{(t)}, \mathbf{y}'_{(t)})$$

where for example $\mathbf{x}_{(t)}$ relates to states of group (1) and has dimension $2L+1$. It then follows that

$$\begin{pmatrix} \mathbf{x}_{(t)} \\ \mathbf{y}_{(t)} \end{pmatrix} = \begin{pmatrix} \mathbf{C} & 2\mathbf{D} \\ \mathbf{E} & 2\mathbf{G} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{(t-1)} \\ \mathbf{y}_{(t-1)} \end{pmatrix}$$

and iteration can be performed with the reduced vector and square matrix of dimension $(2L+1)(L+1)$.

5

Theory of limits to selection with line crossing

by

William G. Hill

From:

Mathematical Topics in Population Genetics

Edited by

Ken-ichi Kojima

With 55 Figures

Springer-Verlag Berlin·Heidelberg·New York 1970

Theory of Limits to Selection with Line Crossing

W. G. Hill

Introduction

Breeders of many species of animals and plants make breed or strain crosses and market the crossbred progeny in order to utilize heterosis. Several breeding schemes have been suggested for the improvement of such crossbreds without necessarily improving the parental strains. (The progeny will be called "crossbred" even if their parents are of different strains but from the same breed). There are essentially two types of programme. In the first, most of the emphasis is based on selection *between* lines. Typically these lines are developed by rapid inbreeding, and little or no selection is practised within them. Large scale testing is then undertaken to find those lines which produce the best cross. Other breeding programmes typically start with a pair of lines already known to give potentially useful crossbreds, and then selection is practised *within* these lines, with the aim of improving this cross. Of course, mixtures of these schemes are also used in practice.

In this paper the analysis will be almost entirely restricted to the alternative methods of practising selection within lines. This may be based only on performance in the individual pure strains, which will be referred to as pure line selection (PLS). Alternatively, selection can be carried out in one line for cross performance against a tester strain which may be highly inbred. This method was proposed by Hull (1945) and is commonly called recurrent selection to a tester (RST). Comstock, Robinson, and Harvey (1949) suggested the method of reciprocal recurrent selection (RRS) in which selection on cross performance is practised within both populations making the cross. However it is not necessary for us to assume that reciprocal crosses between the populations are actually made. In both the RST and RRS methods, individuals are crossed to the other strain, and those which have the best crossbred progeny or half-sibs are selected as parents of the next generation. These methods are thus intended to utilize non-additive genetic variation, particularly from overdominant genes.

Theoretical comparisons of the efficiency of the alternative selection and crossing schemes have been made by Comstock *et al.* (1949), Dicker-

son (1952), Crow (1953), and Cress (1966). However, as Bowman (1959) has pointed out in a review, these theoretical calculations are generally based on three suppositions: (1) no epistasis, (2) no more than two alleles per locus, and (3) linkage equilibrium. Bowman considers that if the literature regarding heterosis is taken into account, then comparisons based on these assumptions must be of limited value. In fact, these studies also make another important assumption, namely that the populations are of very large size (i.e. approaching infinitely large size) so that the maximum improvement possible in any scheme would be expected to be attained eventually. Robertson (1960), however, has drawn attention to the problems of limits to selection in populations of finite size. During the selection programme some favourable genes may be lost from the population by chance, so that the final advance depends on the effective population size as well as the initial frequencies and effects of the individual genes. Also, average rates of advance are likely to be influenced by population size. This is particularly important when genes with heterozygote superiority are initially at equilibrium so that no progress would initially be made in an RRS programme. Both Arthur (1964) and Cress (1967) have used Monte Carlo methods to study the effects of initial restriction of population size in causing drift from equilibrium gene frequency situations. Also Dickerson (1952) showed that RST with a highly inbred tester could be more efficient in terms of initial response than RRS since there is then no unstable equilibrium state. However, in Arthur's model the gene effects, selection intensity, and population size were sufficiently large that eventually the best possible limit was attained, and Cress investigated the rate of advance for only five generations.

Some long-term experimental comparisons of alternative cross breeding experiments have been made with *Drosophila*. Bell, Moore, and Warren (1955) compared several selection schemes for improvement of egg size and production in *D. melanogaster*. Egg size appeared to be largely controlled by additive genes, and conventional pure line family selection proved most efficient. For egg production, Bell *et al.* found that the response with RRS was superior to that with either PLS or RST. However the RRS line was not superior to the best single crosses between inbred lines developed (with much less effort) from the base populations. Rasmuson (1956) compared PLS and RRS for egg production, hatchability, and body weight in *D. melanogaster*. Only with egg production was more progress made with RRS and then by just 6–7%, with most of this superiority obtained after a few generations of selection. Kojima and Kelleher (1963) obtained substantial response to RRS for egg production in *D. pseudoobscura* but only for the first 10 or so generations of selection, after which the population reached a plateau at a

level equivalent to that of the best 4% of all possible two-way crosses in the base population. The earlier experiments are reviewed in more detail by Bowman (1959) who concluded that no direct proof had been published at that time to indicate that the methods of RST and RRS were at all successful in what they were theoretically intended to achieve, namely to utilize non-additive variation. Large experiments with breeding schemes in poultry are being carried out in the United States but the results are as yet, unpublished.

Since the earlier theoretical studies have been primarily concerned with infinitely large populations, or have not considered the finite population model in any detail, it seemed worthwhile to study the expected rates of advance in a finite model with the alternative within-line selection schemes for producing crosses. The theory may help to throw some light on the experimental results and perhaps give some pointers towards breeding practice. Unfortunately there are so many possible parameters which can be studied that a very simple model has to be used, and several approximations will be made in the course of the analysis. In particular, we shall ignore epistasis and linkage and, again in common with Comstock *et al.* (1949) and Dickerson (1952), assume that there are only two alleles per locus. For simplicity, we shall usually also assume that the tester strain in an RST programme is already homozygous, or that it becomes homozygous immediately with inbreeding. With poultry, for example, this idealized situation cannot be attained, and the errors in this approximation will be investigated briefly.

Model. Response to a Single Cycle of Selection

Let us consider an autosomal locus with alternative alleles A_1 and A_2. The average genotypic values of A_1A_1 and A_2A_2 are assumed to be a_1 and a_2 units, respectively, poorer than the heterozygote for the quantitative trait under selection. We let q be the frequency of A_1 and let $\bar{q} = a_2/(a_1 + a_2)$. Since only differences in genotypic value are important, we can arbitrarily let the genotypic value of the heterozygote be $a = a_1 + a_2$. Thus we have:

Genotype	A_1A_1	A_1A_2	A_2A_2
Genotypic value	$(a_1+a_2)-a_1$	a_1+a_2	$(a_1+a_2)-a_2$
$=$	$a\bar{q}$	a	$a(1-\bar{q})$

The alternative types of gene action can be summarised as follows:

Overdominant	$0 < \bar{q} < 1$
A_1 completely dominant over A_2	$\bar{q} = 1$
A_1 partially dominant over A_2	$1 < \bar{q} < \infty$
A_1 recessive to A_2	$\bar{q} = 0$
Additive	$\bar{q} \to \infty$ but $(a\bar{q})$ finite.

This way of expressing the gene effects is most suitable for the case of overdominant gene action, with which we shall be particularly concerned, when $\bar{q}$ is the equilibrium gene frequency for large populations. Otherwise $\bar{q}$ has no obvious interpretation and merely becomes a convenient parameter. With additive gene action, where $\bar{q}$ becomes infinite, $\bar{q}$ always appears in expressions for changes in gene frequency as $a\bar{q}$, which is assumed to be finite. Although this definition of the model is less suitable for additive gene action, we shall rarely be concerned with additivity in a theory of selection for cross performance. In other definitions of gene effects (e.g. Comstock *et al.*, 1949; Dickerson, 1952) some terms become infinitely large when there is "pure" overdominance which is $\bar{q} = 0.5$ in this model.

1. Pure Line Selection

If truncation selection is practised on individual phenotypes in a large single population the change in gene frequency in one generation is, approximately,

$$\delta q = \frac{-i_m a}{\sigma} q(1-q)(q-\bar{q}) \qquad (1)$$

where i_m is the selection differential in standard units and σ is the phenotypic standard deviation. Formulae similar to Eq. (1) have been derived by various authors, notably Haldane (1931), Comstock *et al.* (1949), Crow (1953), and Griffing (1960). Eq. (1) holds only if gene effects are small such that terms in $(a/\sigma)^r$ can be ignored relative to a/σ for $r > 1$. Latter (1965) has examined the errors induced by this approximation for additive gene action in infinite populations. In populations of finite size δq represents the expected (i.e. mean) change in gene frequency, and can be predicted more accurately if i_m is computed from order statistics than from the normal integral. A more detailed discussion is given by Kojima (1961) and Hill (1969a). If progeny testing, for example, is practised in a pure line, the response becomes

$$\delta q = -\tfrac{1}{2} i a/\sigma_f \qquad (2)$$

where σ_f is the standard deviation of progeny test means. More generally the response will be proportional to the average of i/σ_f for the two sexes, if, as is probable, they are not tested with exactly the same design. The relative responses with different schemes, such as individual selection or progeny testing are well known (e.g. Falconer, 1960), and we shall return to the problem of generation interval later. Of course Eq. (2) is subject to the same assumptions as Eq. (1) on population size and magnitude of gene effects. The formulae can be greatly simplified if we let $s = ia/\sigma_f$ so that

$$\delta q = -\tfrac{1}{2} sq(1-q)(q-\bar{q}). \tag{3}$$

Thus s may be regarded as a selective value.

2. Selection on Cross Performance

In an RRS scheme those individuals with the highest average cross-bred progeny test are assumed to be chosen as parents of the next generation. Let us denote the populations X and Y, and for the allele A_1 let p and q be their respective frequencies and r and s their respective selective values (i.e. the mean over sexes of ia/σ_f). Predictions of changes in gene frequency in RRS programmes have been given by Comstock *et al.* (1949) and Dickerson (1952). They are similar to those for pure line selection and will be stated here without derivation. With small effects and progeny testing the changes in gene frequency will be:

$$\begin{aligned} &\text{Population } X\text{: } \delta p = -\tfrac{1}{2} rp(1-p)(q-\bar{q}), \\ &\text{Population } Y\text{: } \delta q = -\tfrac{1}{2} sq(1-q)(p-\bar{q}). \end{aligned} \tag{4}$$

In an RRS programme we might expect r and s to be equal, but this would not be the case if no reciprocal crosses were made.

If the tester strain, X, is assumed to be homozygous in an RST breeding programme, the response in population Y will depend on which allele is fixed in X. With a similar progeny testing scheme as in the RRS programme, changes in gene frequency will be as follows:

Allele fixed in population X	Changes in gene frequency in Y	
$A_1 (p=1)$	$\delta q = -\tfrac{1}{2} sq(1-q)(1-\bar{q})$	(5a)
$A_2 (p=0)$	$\delta q = \tfrac{1}{2} sq(1-q)\bar{q}$.	(5b)

In general, of course, Eq. (4) can be applied to RST also, even if the tester is not inbred.

When comparing the PLS, RRS, and RST schemes we may not assume that the selective value, s, is the same in each case, because the selection intensities or variances of progeny test means may not be the same. For example, σ_f may be less in an RST scheme if the tester is homozygous.

3. Changes in the Mean of the Quantitative Trait

If random mating is practised between individuals of the opposite strain, the mean, μ, of the crossbred progeny for the quantitative trait is

$$\mu = a[1 - \bar{q}(1 - \bar{q}) - (p - \bar{q})(q - \bar{q})] . \quad (6)$$

This mean is maximized with overdominance if $p = 1$ and $q = 0$ or vice versa, with complete dominance if $p = 1$ or $q = 1$, and with partial dominance or additivity if $p = q = 1$. The change in the mean with one cycle of selection is:

$$\delta\mu = -a[(q - \bar{q})\,\delta p + (p - \bar{q})\,\delta q + \delta p \delta q] .$$

Thus if the product term $\delta p \delta q$ is ignored, which should introduce little error if changes in gene frequency are small each generation, the responses to a single cycle of selection for the alternative schemes are as follows:

System		$\delta\mu$
PLS		$\frac{a}{2}(p - \bar{q})(q - \bar{q})[rp(1-p) + sq(1-q)]$
RRS		$\frac{a}{2}[rp(1-p)(q - \bar{q})^2 + sq(1-q)(p - \bar{q})^2]$
RST	$p = 1$	$\frac{a}{2} sq(1-q)(1 - \bar{q})^2$
	$p = 0$	$\frac{a}{2} sq(1-q)\bar{q}^2$

In the PLS system selection is carried out independently in the two populations, which are then crossed. The selective values are assumed to be r and s in populations X and Y, respectively.

Extension of the Theory to Finite Populations

Explanation of the theory for finite populations may be clarified if we first consider selection in a single population, for which the theory of limits was first discussed by Robertson (1960). Imagine that an identical selection programme is practised in a large number of replicate lines in which the frequency of the A_1 allele was originally q_0. With genetic sampling (drift) the gene frequencies will no longer remain the same in all lines, and we can thus discuss the distribution of gene frequency among these lines. Eventually all lines will reach fixation and a limit will be reached, although with overdominant genes fixation may occur very slowly, and accompany a decline in response (Robertson, 1962; Hill and Robertson, 1968). The chance of fixation of the allele A_1 is defined as the proportion of lines in which it is eventually fixed, for specified initial frequency.

Changes in the mean of a cross between two finite populations have to be studied in terms of the joint distribution of gene frequency and joint probabilities of fixation in the two lines. Thus we let $w(p_0, q_0)$, or simply w, be the probability that A_1 is fixed in both lines X and Y, given that their initial frequencies were p_0 and q_0 respectively. Similarly we define $u(p_0)$ and $v(q_0)$, or u and v, as the marginal probabilities of fixation of A_1 in X and Y respectively. Drift occurs independently in the two populations, but changes in gene frequency with selection are not independent in an RRS programme. With pure line selection we can assume that $w = uv$, but in a successful reciprocal recurrent selection programme for an overdominant gene we would hope that A_1 was not frequently fixed in both lines making the cross, so that $w < uv$. The probabilities of fixation in the various states can be summarized thus:

Allele fixed in X, Y	A_1, A_1	A_1, A_2	A_2, A_1	A_2, A_2
Probability	w	$u-w$	$v-w$	$1-u-v+w$
Crossbred mean	$a\bar{q}$	a	a	$a(1-\bar{q})$

Let us assume that the effective sizes of the populations X and Y are M and N respectively. With pure line selection in population Y, for example, the conditional probability that among the total of $2N$ alleles at some cycle $t+1$ there are jA_1 alleles, given that there were iA_1 alleles at generation t, can be shown to be approximately

$$b_{ij} = \binom{2N}{j}(q+\delta q)^j(1-q-\delta q)^{2N-j}, \quad 0 \leqq i, j \leqq 2N \tag{7}$$

where $q = i/2N$, the gene frequency at generation t, and $\delta q = -\frac{1}{2}sq(1-q)(q-q)$ from Eq. (3). The approximations associated with Eq. (7) for truncation selection are discussed by Hill (1969a), but in the context of individual selection rather than progeny testing. With PLS b_{ij} is independent of t if it is assumed that the selective values do not change, and similar transition probabilities can be specified for selection in X. Similarly, with RST, so long as the tester strain is homozygous, δq can be evaluated using Eq. (5a) or (5b) and the transition probabilities of the form of Eq. (7) become independent of t. But with RRS, although single generation responses can be expressed in this form, it is necessary to consider the joint transition probabilities in the two populations in order to describe long-term response. Since the genetic sampling occurs independently in the two populations, we have

$$d_{(h,i,j,k)} = \binom{2M}{j}(p+\delta p)^j(1-p-\delta p)^{2M-j}\binom{2N}{k}(q+\delta q)^k(1-q-\delta q)^{2N-k} \tag{8}$$

where $p = h/2M$; $q = i/2N$; $0 \leqq h, j \leqq 2M$; $0 \leqq i, k \leqq 2N$;

$$\delta p = -\frac{r}{2} p(1-p)(q-\bar{q}); \quad \delta q = -\frac{s}{2} q(1-q)(p-\bar{q})$$

and $d_{(h,i,j,k)}$ is the probability that, conditional on there being h and i A_1 alleles in populations X and Y respectively at some cycle t, there are j and k respectively in the succeeding cycle. Again, so long as r and s remain constant for all t, so does $d_{(h,i,j,k)}$. Of course transition probabilities for PLS can be written in the same form as Eq. (8), but they factor into terms of the form of Eq. (7).

The transition probabilities of Eq. (7) and Eq. (8) can be expressed in matrix form. In Eq. (8) a row of the matrix specifies both h and i and a column j and k so that the matrix is square of dimension $(2M+1)(2N+1)$. Standard techniques can then be used to obtain numerical results for distribution of gene frequency and the chances of fixation, which need not be discussed here. The type of method adopted is described elsewhere (Hill, 1969a). On an I.C.T. Atlas computer it was practicable to work with $N = M = 8$ with the transition matrix for RRS.

Diffusion Equation and Simple Approximations

Diffusion models have been widely applied to problems of selection at a single locus in a single finite population (Kimura, 1964). In particular, the chance of fixation has been derived for such models (Kimura, 1957) and has been applied to the theory of selection limits (Robertson, 1960). The diffusion equation is continuous in both time and gene frequency, but the chance of fixation computed from the equation has been shown to be a good predictor for a model of artificial selection in a finite population with discrete generations and values of gene frequency (Hill, 1969a).

1. Pure Line Selection

If the diffusion equation is used to approximate the gene frequency distribution for selection in a single finite population, say Y, and if time is measured on a scale proportional to the effective population size N the selection advance is a function of only Ns, $\bar{q}$, and the initial frequency, q_0 (Robertson, 1962). If the expected change in gene frequency is given by Eq. (3), the chance of fixation of A_1 in Y is given

$$v = \int_0^{q_0} e^{Ns(x-\bar{q})^2} dx \Big/ \int_0^1 e^{Ns(x-\bar{q})^2} dx \qquad (9)$$

(Kimura, 1957) and similarly u is given in terms of Nr, $\bar{q}$, and p_0 in population X. As we have mentioned, the joint chance of fixation of A_1 in the two populations is the product of the marginal probabilities, so that μ_L, the crossbred mean at the limit, is given by substitution in Eq. (6) as

$$\mu_L = a[1 - \bar{q}(1 - \bar{q}) - (u - \bar{q})(v - \bar{q})] . \tag{10}$$

If Ns and $Ns\bar{q}$ are small relative to unity, series expansion of Eq. (9) yields

$$v = q_0 + Nsq_0(1 - q_0)[\bar{q} - \tfrac{1}{3}(1 + q_0)] + 0[(Ns)^2] . \tag{11}$$

Robertson (1960) has expressed v in this form for $\bar{q} = 1$.

Under the diffusion approximations the selection limit, μ_L, will be a function of Mr, Ns, p_0, q_0, and $\bar{q}$. We shall be particularly interested in the case where the same breeding programme is practised in each line, so that $M = N$ and $Mr = Ns$. The time scale of the process will then be proportional to N.

2. Recurrent Selection to an Inbred Tester

If the tester strain X is homozygous at locus A, selection in line Y for cross performance is then equivalent to selection for an additive gene with selective value $-s(1 - \bar{q})/2$ or $s\bar{q}/2$ depending on whether A_1 or A_2, respectively, is fixed in X (Eq. (4)). Thus, from Kimura (1957), the chance of fixation of A_1 in Y is given by

$$v = [1 - e^{2Ns(p^* - \bar{q})q_0}]/[1 - e^{2Ns(p^* - \bar{q})}] \tag{12}$$

where p^* is the frequency of A_1 in the tester and takes values $p^* = 0$ or 1. If X were a cross between two homozygous lines, p^* could then take the value 0.5. The population mean at the limit is simply obtained from the definition of genotypic values.

If population X initially is segregating at locus A, but is inbred very rapidly without selection so that we can assume it is homozygous from the outset of the RST programme, the probability that X is fixed for A_1 is p_0 and the probability is $1 - p_0$ that X is fixed for A_2. The expected value of the selection limit becomes

$$\mu_L = a[1 - (1 - \bar{q})p_0 v_1 - \bar{q}(1 - p_0)v_2] \tag{13}$$

where v_1, v_2 are the conditional chances of fixation of A_1 in Y given that A_1, A_2 are fixed in X and are given by substitution into Eq. (12) of $p^* = 1$ or $p^* = 0$, respectively.

If Ns and $Ns\bar{q}$ are of small order, Eq. (12) and (13) become

$$v = q_0 + Nsq_0(1 - q_0)(p^* - \bar{q}) + 0[(Ns)^2] , \tag{14}$$

$$\mu_L - \mu_0 = Nsaq_0(1 - q_0)[p_0(1 - p_0) + (p_0 - \bar{q})^2] . \tag{15}$$

3. Reciprocal Recurrent Selection

No explicit formulae for the limit have been obtained from diffusion equations for RRS. However, a formula developed by Kimura (personal communication) and cited by Ohta (1968) for a special case of additive x additive epistatic interaction between independent loci in a single population can be modified to give the selection limit for RRS with complete dominance ($\bar{q}=1$) when the same breeding programme is practised in both lines, i.e. $N=M, r=s$. The selection limit, μ_L, in the cross becomes

$$\mu_L = a[e^{2Ns} - e^{2Ns(1-p_0)(1-q_0)}]/(e^{2Ns}-1). \tag{16}$$

However, both forward and backward Kolmogorov diffusion equations can be set up to approximate selection with RRS for all values of $\bar{q}$ and $N \neq M, r \neq s$. The equations can be obtained by substitution into the multivariate formulae given by Kimura (1964), and we make no attempt at rigour here. Let $\phi(p, q, p_0, q_0, t)$ be the joint density of the gene frequency distribution in the two populations at time t, for initial frequencies p_0 and q_0. The mean changes in gene frequency ($M_{\delta p}$ and $M_{\delta q}$ of Kimura) are given by Eq. (4), the variances of changes ($V_{\delta p}$ and $V_{\delta q}$) are

$$V_{\delta p} = \frac{p(1-p)}{2M}, \qquad V_{\delta q} = \frac{q(1-q)}{2N}$$

and the covariance of change is zero since sampling of genes occurs independently in the two populations. The forward equation is

$$\begin{aligned}\frac{\partial \phi}{\partial t} = {} & \frac{1}{2}\frac{\partial^2}{\partial p^2}\left[\frac{p(1-p)}{2M}\phi\right] + \frac{1}{2}\frac{\partial^2}{\partial q^2}\left[\frac{q(1-q)}{2N}\phi\right] \\ & + \frac{\partial}{\partial p}\left[\frac{1}{2}rp(1-p)(q-\bar{q})\phi\right] + \frac{\partial}{\partial q}\left[\frac{1}{2}sq(1-q)(p-\bar{q})\phi\right].\end{aligned} \tag{17}$$

We are particularly interested in the model in which the same breeding programme is practised in both populations (with $M=N$ and $r=s$). Eq. (17) may be written

$$\begin{aligned}\frac{\partial \phi}{\partial (t/N)} = {} & \frac{1}{4}\frac{\partial^2}{\partial p^2}[p(1-p)\phi] + \frac{1}{4}\frac{\partial^2}{\partial q^2}[q(1-q)\phi] \\ & + \frac{1}{2}Ns\left\{\frac{\partial}{\partial p}[p(1-p)(q-\bar{q})\phi] + \frac{\partial}{\partial q}[q(1-q)(p-\bar{q})\phi]\right\}.\end{aligned} \tag{18}$$

Therefore on a time scale proportional to N, and under the diffusion equation assumptions, changes in gene frequency will be a function of only Ns, $\bar{q}$, and the initial frequencies p_0 and q_0.

When r and s are very small, the value of μ_L can be obtained approximately for RRS for all values of $\bar{q}$ and $N \neq M$, $r \neq s$, and the derivation is given in the appendix. The main result is, from Eq. (A3) of the appendix

$$\mu_L - \mu_0 = a\left\{Mrp_0(1-p_0)(q_0-\bar{q})^2 + Nsq_0(1-q_0)(p_0-\bar{q})^2 + \left[Mr+Ns-\frac{2MN(r+s)}{2M+2N-1}\right]p_0(1-p_0)q_0(1-q_0)\right\} \quad (19)$$

plus terms containing higher powers of r and s. With $M=N$ and $r=s$, Eq. (19) reduces to

$$\mu_L - \mu_0 = Nsa\left[p_0(1-p_0)(q_0-\bar{q})^2 + q_0(1-q_0)(p_0-\bar{q})^2 + \left(1+\frac{1}{4N-1}\right)p_0(1-p_0)q_0(1-q_0)\right]. \quad (20)$$

If N is large, the term $1/(N-1)$ in Eq. (20) may be ignored, and when $\bar{q}=1$ (complete dominance), Eq. (20) reduces to

$$\mu_L - \mu_0 = Nsa(1-p_0)(1-q_0)(p_0+q_0-p_0q_0) \quad (21)$$

which can also be obtained by series expansion of Eq. (16), taking only terms up to order Ns.

Checks on the adequacy of the diffusion equation approximation for chances of fixation in single populations have been reported (Ewens, 1963; Hill, 1969a). These indicate that the diffusion results are certainly adequate for descriptive purposes and can therefore be used for PLS and RST in this study. In Table 1 diffusion equation results for RRS with complete dominance (Eq. (16)) are compared with those from the transition probability matrix (8) for $N=4$ and $N=8$. We see that the

Table 1. *Comparison of selection limits μ_L/a for reciprocal recurrent selection with complete dominance estimated from transition probability matrices with $N=4$(T4) and $N=8$(T8), from the diffusion approximation (DA), and by approximation ignoring terms of order greater than Ns (A)*

	Initial frequencies											
	0.25, 0.25				0.25, 0.5				0.5, 0.5			
Ns	T4	T8	DA	A	T4	T8	DA	A	T4	T8	DA	A
$\frac{1}{2}$	0.5589	0.5598	0.5606	0.5605	0.7336	0.7345	0.7352	0.7422	0.8329	0.8339	0.8347	0.8437
1	0.6702	0.6723	0.6744	0.6836	0.8223	0.8237	0.8252	0.8594	0.8956	0.8971	0.8985	0.9375
2	0.8329	0.8370	0.8416	—	0.9311	0.9329	0.9350	—	0.9650	0.9664	0.9679	—
4	0.9618	0.9655	0.9701	—	0.9918	0.9925	0.9936	—	0.9970	0.9974	0.9979	—
8	0.9977	0.9983	0.9991	—	0.9999	0.9999	1.0000	—	1.0000	1.0000	1.0000	—

agreement is generally very good, although for a given value of Ns the limit is consistently higher with larger N and with the diffusion approximation, which may be regarded as the limiting value as N becomes infinite. A similar bias has been observed in single population studies. Comparisons of alternative breeding methods have therefore always been made at the *same* population size, either using the diffusion equation or transition probability matrices with $N=8$ for RRS, RST, and PLS. However, the results of Table 1 indicate that we can express the parameters in terms of Ns rather than N and s separately and thus draw inferences about a wide range of population sizes from results obtained at one population size, thereby reducing greatly the amount of computation required.

Also included in Table 1 is a check on the accuracy of the simple formula (21) including only linear terms in Ns. For $Ns=\frac{1}{2}$ there seems to be reasonable agreement, but the simple formula becomes strongly biased upwards if Ns is much larger. This range of validity is not surprising, since μ_L can be written as an expansion of exponential terms in Ns from Eq. (16).

Comparison of Selection Methods

We now have sufficient theory to enable us to compare the efficiencies of the alternative breeding schemes. Since the relative efficiencies differ considerably from one model of gene action to another, we shall consider these in turn, starting with complete dominance. All comparisons will be made in terms of the parameters Mr, Ns, and so on, although the cases with $Mr=Ns$ will be studied in greatest detail.

In practice it should be possible to attain higher values of s and Ns with pure line selection in particular, for recourse need not be made to progeny testing and the generation interval can be reduced. Using as a basis a single two year cycle of one generation with progeny testing, we have defined $s=\frac{ia}{\sigma_f}$ where σ_f is the standard deviation of progeny test means, and the response *per cycle* with PLS is $\delta q=-\frac{s}{2}q(1-q)(q-\bar{q})$. With mass selection (and, of course, PLS) the response *per generation* (Eq. (1)) is $\delta q=-\frac{i_m a}{\sigma}q(1-q)(q-\bar{q})$ where σ is the phenotypic standard deviation. Thus the approximate value of s over two generations is $\frac{4i_m a}{\sigma}$, but the effective population size for the two generation period is $N/2$ since the sampling variance for one generation with $N/2$ is approxi-

mately the same as with two generations with population size N. An example may help to illustrate these comparisons.

(i) Progeny testing: two year cycle $\sigma_f = 8$, $i = 1$, $N = 32$.

(ii) Mass selection: one year cycle $\sigma = 30$, $i_m = 1.5$, population size = 40 each generation. Also let $a = 2$, with units, say, eggs per hen in poultry. With progeny testing we have $N = 32$, $s = 1/4$, $Ns = 8$ and in the same context with mass selection $N = 20$, $s = 2/5$, $Ns = 8$. We might thus predict the same total advance using these two schemes with pure line selection, since the limit is a function of Ns, but the rate of advance, inversely proportional to N, would be faster with mass selection. Other variations on selection programmes which affect their efficiency can be included in the same way. Comparisons of the different schemes should not necessarily be made for the same values of Ns; these Ns values can be modified to suit the individual's own prejudices. Unfortunately in the simple model which has been adopted, it is not possible to modify the values of s as selection proceeds, when, for example, some loci may become fixed and the phenotypic variance is reduced. Strictly, therefore, we are considering only single genes.

1. Complete Dominance ($\bar{q} = 1$)

With complete dominance the optimum selection limit is attained if either population is fixed for the favourable dominant allele A_1 and the optimum crossbred can therefore be reached with RST.

In Fig. 1 the selection limit, μ_L, is plotted for the three alternative breeding schemes PLS, RRS, and RST for a model of complete dominance and three different pairs of initial frequencies. The mean at the limit is plotted relative to the gene effect and so it can range from 0 to 1. The results were obtained from the diffusion equation and $Mr = Ns$. With RST it is assumed that selection is practised in one population and the tester is fixed instantaneously, with probability p_0, say, that it is fixed for A_1. When $p_0 \neq q_0$, two alternative sets of results are shown for RST, depending on the initial frequency in the population used as the tester.

For a given value of Ns it can be seen that the highest limit is always attained with RRS, but the superiority over PLS is never large. A doubling of s using PLS, which might be attained by practising mass selection thus reducing the generation interval and increasing the selection intensity, would reverse these rankings. The RST method is competitive so long as selection is practised in the correct population (an ideal which might be difficult to attain in a breeding programme!). We notice in Fig. 1 that the greater response occurs if the selected line has the higher initial frequency, in this case 0.5 as opposed to 0.1. It is coincidental in this

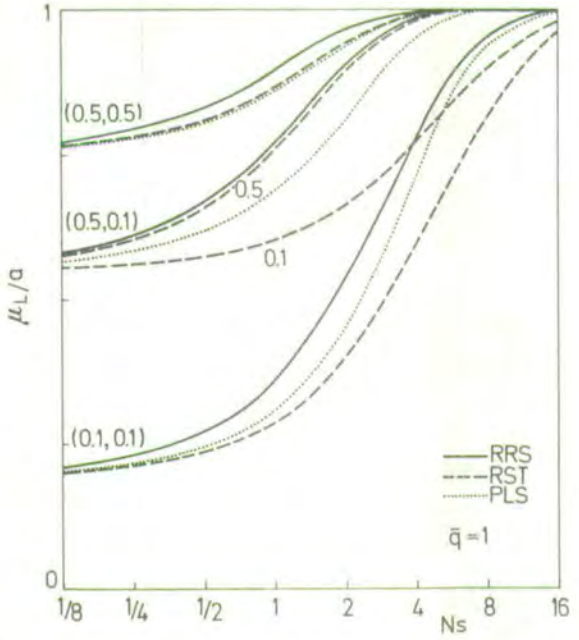

Fig. 1. The selection limit for the crossbred mean with complete dominance expressed as proportion of the gene effect, a, for three pairs of initial frequencies and a range of Ns values. The initial frequency of the selected population with RST is also shown

example that the frequency is also nearer 0.5 where the variance of gene frequency is highest for we now show that it is always more efficient to select by RST in the line at higher initial frequency if there is complete dominance. From Eq. (12) and assuming selection in population Y, with initial frequency q_0, we find with RST that if the tester is fixed instantaneously

$$\mu_L = a[p_0 + (1-p_0)(1-e^{-2Nsq_0})/(1-e^{-2Ns})]\,. \tag{22}$$

With rearrangement and series expansion Eq. (22) gives

$$\mu_L/a = 1 - \frac{2Ns(1-p_0)(1-q_0)}{e^{2Ns}-1}\left\{1 + \frac{2Ns(1-q_0)}{2!} + \frac{[2Ns(1-q_0)]^2}{3!} + \cdots\right\}. \tag{23}$$

For $s>0$ a higher limit is therefore reached if $q_0 > p_0$ than if $p_0 > q_0$, so that selection should be practised in the population with higher initial frequency. The most plausible verbal interpretation seems to be that by selection in the population at high frequency we almost ensure fixation of the favourable allele. Since this is dominant, the optimum limit is reached in the cross.

In the RST system only one population has to be maintained at a large size (i.e. N), for drift is of no consequence in the tester. Thus if the total number of breeding animals is restricted by the facilities available, it may be possible to use a value of N almost twice as large in the single selected population with RST as compared with RRS or PLS, so that the RST system may be relatively more efficient than Fig. 1 indicates. Some of this advantage may be lost if the reproductive rate in the inbred tester is poor, so that family sizes are smaller and the progeny tests less efficient.

When Ns is small, the selection limits under the different systems can easily be compared. The results can be summarised as follows for complete dominance:

$$\begin{aligned} &\text{PLS: } \mu_L - \mu_0 = \tfrac{1}{3} Nsa(1-p_0)(1-q_0)[p_0(2-p_0)+q_0(2-q_0)], \\ &\text{RRS: } \mu_L - \mu_0 = Nsa(1-p_0)(1-q_0)(p_0+q_0-p_0q_0), \\ &\text{RST: } \mu_L - \mu_0 = Nsa(1-p_0)(1-q_0)q_0. \end{aligned} \tag{24}$$

For example if $p_0 \doteq q_0$, the ratios of advance with the different systems are for small Ns, $\dfrac{\text{PLS}}{\text{RST}} = \dfrac{2}{3}(2-q_0)$, $\dfrac{\text{RRS}}{\text{RST}} = 2-q_0$, and $\dfrac{\text{PLS}}{\text{RRS}} = \dfrac{2}{3}$,

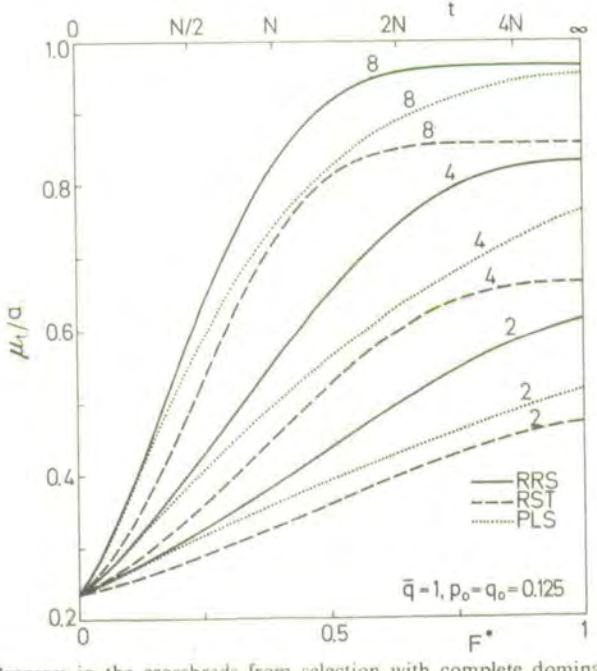

Fig. 2. Progress in the crossbreds from selection with complete dominance, initial frequencies 0.125 in each population and $Ns = 2$, 4 or 8. Time is expressed as $F^* = 1 - e^{-t/2N}$ and the crossbred mean as a proportion of the effect a

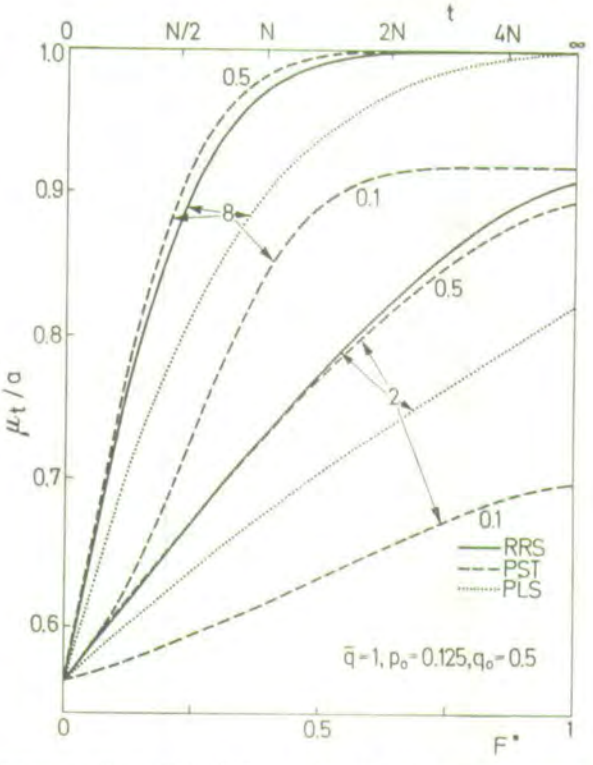

Fig. 3. Progress in the crossbreds from selection with complete dominance, initial frequencies 0.125 and 0.5 and $Ns = 2$ or 8. The initial frequency of the selected population with RST is shown. Time is expressed as $F^* = 1 - e^{t/2N}$ and the crossbred mean as a proportion of the effect a

again assuming that Ns is the same for each system. A doubling of N with RST would yield $\frac{\text{RRS}}{\text{RST}} = 1 - q_0/2$ which is less than 1. Again we note that RST is more efficient at high initial frequencies in the tester. The difference in advance (RRS − PLS) with $p_0 \neq q_0$ may be written $\text{RRS} - \text{PLS} = \frac{1}{3} Nsa(1-p_0)\,(1-q_0)\,[(p_0+q_0) - 5p_0q_0]$ which is positive for $Ns > 0$ and $0 < p_0, q_0 < 1$.

The progress from selection before the limit is reached is shown in Figs. 2 and 3 for two pairs of inital frequencies, $p_0 = q_0 = 0.125$ (Fig. 2) and $p_0 = 0.125$, $q_0 = 0.5$ (Fig. 3), each with $Mr = Ns = 2$ and 8. In these graphs the expected value of μ_t/a at cycle t is plotted in successive cycles. However, the time scale (which ranges to $t \to \infty$) has been condensed by using the transformation $F^* = 1 - e^{-t/2N}$. With no selection F^* is approximately equal to the inbreeding coefficient, where N is the effective population size for a complete cycle of testing and selection. We see in the

figures that although the pattern of response is not identical for the different methods there is essentially no change in ranking from the outset. The selection advance is seen to slow down earlier with RRS and RST than with PLS. Since RST is equivalent to selection for an additive gene, RST and PLS merely reflect the patterns of response for additive and dominant genes in single populations and these have been described elsewhere (Robertson, 1960; Hill, 1969b). Assuming instantaneous fixation of population X with RST, the initial rates of advance ($\delta\mu$ with $t=1$) are

$$\begin{aligned}
&\text{PLS: } \delta\mu = \frac{sa}{2}(1-p_0)(1-q_0)[p_0(1-p_0)+q_0(1-q_0)]\,,\\
&\text{RRS: } \delta\mu = \frac{sa}{2}(1-p_0)(1-q_0)[p_0(1-q_0)+q_0(1-p_0)]\,, \qquad (25)\\
&\text{RST: } \delta\mu = \frac{sa}{2}(1-p_0)(1-q_0)q_0\,.
\end{aligned}$$

Eq. (25) can be compared with the approximate limit formulae (24). The total advance for RST is $2N$ times that in the first generation and rather more than $2N$ times for RRS. However Eq. (24) have very strong restrictions on the magnitude of Ns. Usually less than $2N$ times the initial advance is made with additive gene action (Robertson, 1960; Hill, 1969b).

2. Partial Dominance ($1<\bar{q}<\infty$)

We shall consider the model of partial dominance in less detail than that of either complete dominance or overdominance. In Fig. 4 the crossbred mean at the limit is plotted for the case of $\bar{q}=1.5$. The genotypic values would then be 1.5 a, a and $-0.5\,a$ for A_1A_1, A_1A_2, and A_2A_2 respectively, but have been transformed in the graph to 1.0, 0.75, and 0.0, respectively. Thus if y is the ordinate, $y=0.25+\mu_L/(2a)$ with $0\leqq y\leqq 1$.

The initial frequencies chosen are similar to those given in Fig. 1 for complete dominance, but we notice some distinct differences in the results. Since the optimum crossbred mean is only attained when both populations are fixed for the favoured allele A_1, this limit cannot be reached with RST unless the tester is already fixed for A_1. Thus the RST system is relatively inefficient, particularly when Ns values are large. There is no simple rule about the gene frequencies in the population in which selection should be practised with RST. For $\bar{q}=1.5$, we have from Eq. (12) for RST.

$$\mu_L/a = \frac{3}{2}p_0 - \frac{1}{2} + \frac{1}{2}p_0\left[\frac{1-e^{-Nsq_0}}{1-e^{-Ns}}\right] + \frac{3}{2}(1-p_0)\left[\frac{1-e^{-3Nsq_0}}{1-e^{-3Ns}}\right]. \qquad (26)$$

As $Ns \to \infty$, $\mu_L/a \to 1 + p_0/2$, and the limit is maximized if selection is practised in the population at lower initial frequency, which is capable of making the greatest total advance. For very small Ns, $(\mu_L - \mu_0)/a \to Nsq_0(1-q_0)\ (9/4 - 2q_0)$ which turns out to be higher if q_0 (the initial frequency in the selected population) exceeds p_0, except when q_0 is very close to 0 or 1, when the ranking may be reversed.

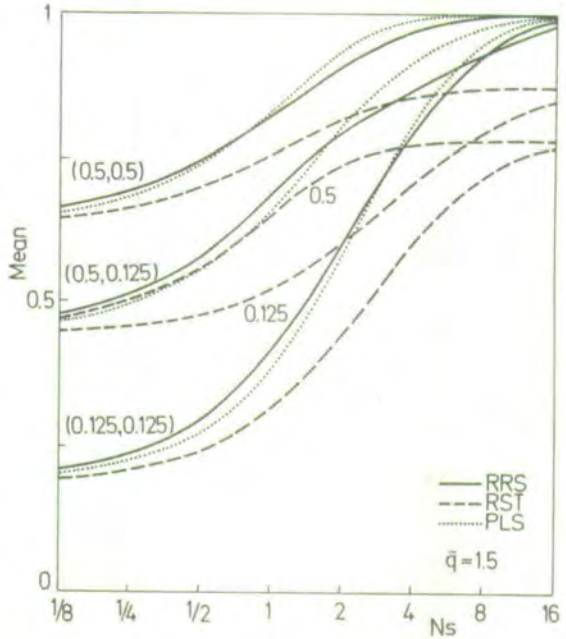

Fig. 4. As Fig. 1, but for partial dominance with $\bar{q} = 1.5$. The mean is expressed as $0.25 + \mu_L/2a$

At low Ns values RRS is more efficient than PLS, but the ranking changes as Ns becomes large. However, at no Ns value is the difference in gain large between these alternative schemes. Since higher Ns values are likely to be possible with PLS, this method is likely to be most efficient in practice if there is partially dominant gene action. We also note in Fig. 4 that with the highest Ns value (16) and RRS a higher limit is attained with $p_0, q_0 = 0.125, 0.5$ than with $p_0 = q_0 = 0.5$. Perhaps the explanation for this phenomenon is that the selection pressures are initially weaker in, for example, population X if Y has frequency $q_0 = 0.5$ rather than 0.125 and X becomes fixed for A_2 in the early generations more frequently when $q_0 = 0.5$.

The rates of advance with partial dominance are similar for RRS and PLS, and we shall not discuss them further.

3. Additivity ($\bar{q} \to \infty$)

If there is additive gene action the ranking of individuals on pure or cross performance will be the same, so that for given Ns both RRS and PLS have the same efficiency and rates of advance, but higher Ns values may be attainable with PLS. The RST system is not suitable for additive gene action, as Comstock *et al.* (1949) and Dickerson (1952) have pointed out, for with selection in only one population the advance can not exceed one-half of that possible with fixation of the favourable allele in both populations.

4. Overdominance ($0 < \bar{q} < 1$)

With overdominant gene action PLS is not a successful breeding system. If both homozygotes have the same genotypic value ($\bar{q} = 0.5$) PLS leads eventually to random fixation of either homozygote. Otherwise the more favourable homozygote is more frequently fixed (Robertson, 1962; Hill and Robertson, 1968), and, especially when $\bar{q}$ lies far from 0.5 and Ns is large, almost all lines will be fixed for the same homozygote and the line cross will not show heterosis (Robertson, 1962). However, with intermediate equilibrium frequencies and large Ns the progress to fixation is very slow, and we might expect to find populations selected on pure line performance having overdominant genes near their equilibrium frequency. This situation has received some attention, for the initial rate of advance with RRS will be zero if both populations are at equilibrium (Comstock *et al.*, 1949; Dickerson, 1952). Arthur (1964) demonstrated that the early response with RRS is greatly increased by a short period of inbreeding before selection is started, so that the genes drift from their equilibrium frequency. Dickerson (1952) showed that recurrent selection to an inbred tester gave greater initial response, for additive variance is immediately obtained. Much of the following discussion on the overdominance model will therefore center on the case of initial equilibrium; PLS will be ignored.

Initial Equilibrium. The mean of either line or the crossbred is initially $\mu_0 = a[1 - \bar{q}(1 - \bar{q})]$. If the maximum possible gain is made, the final crosses will all be heterozygotes with mean a, and this gain is $a\bar{q}(1 - \bar{q})$. It also turns out that at equilibrium $\sigma_d = a\bar{q}(1 - \bar{q})$, where σ_d^2 is the dominance variance at this locus. There is, of course, no additive variance at equilibrium. The initial rate of advance is $\delta\mu = 0$ with RRS and $\delta\mu = \frac{sa}{2}\bar{q}(1 - \bar{q})^3$ or $\delta\mu = \frac{sa}{2}\bar{q}^3(1 - \bar{q})$ with RST, according to whether A_1 or A_2, respectively, is fixed in the inbred tester. As before we

shall assume that one population is fixed instantaneously, with probability $\bar{q}$ that A_1 is the allele fixed. Then the initial rate of advance with RST is on average $\delta\mu = \frac{sa}{2}\bar{q}^2(1-\bar{q})^2 = i\sigma_d^2/2\sigma_f$. The additive variance in the cross is now σ_d^2. Only after some drift has occurred is any additive variance generated with RRS and a response obtained.

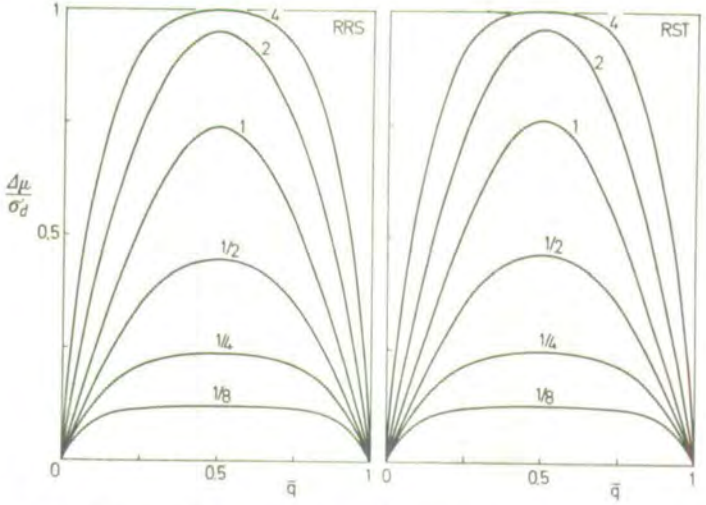

Fig. 5. Ratio of expected to maximum gain ($\Delta\mu/\sigma_d$) in the crossbreds for overdominance with initial equilibrium. Curves are plotted for different values of $Ni\sigma_d/\sigma_f$

In Fig. 5 the selection limits with initial equilibrium for RRS and RST (assuming instantaneous fixation of the tester) are compared for all values of $\bar{q}$, using matrix iteration results with $N = 8$. Since the total advance $\Delta\mu = \mu_L - \mu_0$ lies in the range $0 \leqq \Delta\mu \leqq \sigma_d$, for all $\bar{q}$, the quantity shown, $\Delta\mu/\sigma_d$, is the proportion of the possible gain which is realized. Using the diffusion equation approximation we know that, if $p_0 = q_0 = \bar{q}$ and $Mr = Ns$, $\Delta\mu$ is a function of only $Ns = Nia/\sigma_f$ and $\bar{q}$. Therefore $\Delta\mu$ can also be expressed as a function of only $Nia\bar{q}(1-\bar{q})/\sigma_f = Ni\sigma_d/\sigma_f$ and $\bar{q}$, where Ni is under the breeder's control and σ_d/σ_f is a measure of the contribution of the locus in question to the total variability. So in Fig. 5 the limit is expressed as $\Delta\mu/\sigma_d$ for a range of values of $Ni\sigma_d/\sigma_f$. Clearly the most startling aspect of the results is that almost the same advance is made with RST as with RRS for the whole spectrum of parameters. For all but the smallest Ns values an algebraic verification has not been found, since we have no diffusion formula for the limit with RRS. However when second order terms in r and s are ignored, we have

from Eq. (19) that with initial equilibrium

$$\mu_L - \mu_0 = a\left[Mr + Ns - \frac{2MN(r+s)}{2M+2N-1}\right]\bar{q}^2(1-\bar{q})^2 . \qquad (27)$$

For $Mr = Ns$, ignoring terms of order $1/N$ relative to 1 and setting $Ns\ a\ \bar{q}^2(1-\bar{q})^2 = Ni\sigma_d^2/\sigma_f$ Eq. (19) reduces to

$$\mu_L - \mu_0 = Ni\sigma_d^2/\sigma_f . \qquad (28)$$

Summing over all overdominant loci at equilibrium, let $\sigma_D^2 = \Sigma\sigma_d^2$ so that the total advance with RRS from equilibrium becomes $Ni\sigma_D^2/\sigma_f$, assuming gene effects are small. Setting $Mr = 0$ in Eq. (20) we obtain the predicted total advance with RST when the tester is fixed instantaneously, which is again $\mu_L - \mu_0 = Ni\sigma_D^2/\sigma_f$. This formula for RST can also be obtained directly from the results of Robertson (1960), who showed that the total advance with selection for additive genes of small effect is $2Ni_m\sigma_A^2/\sigma$, where σ_A^2 is the additive variance and mass selection is practised. With progeny testing, Robertson's formula becomes $Ni\sigma_A^2/\sigma_f$ in the notation of this paper.

Therefore we see that RRS and RST are predicted to give the same total advance if gene effects are small. This advance, $Ni\sigma_D^2/\sigma_f$, is independent of $\bar{q}$, and similarly at a single locus the advance $\Delta\mu$, expressed as a proportion of the possible advance, $\Delta\mu/\sigma_d$, is $Ni\sigma_d/\sigma_f$, again independent of $\bar{q}$. We see in Fig. 5 that, for given $Ni\sigma_d/\sigma_f$, this relation does not hold well if the $\bar{q}$ values are extreme. This is not surprising, since larger values of ia/σ_f (selective values) are required for the same value of $i\sigma_d/\sigma_f$, and the assumptions of small effects are violated. Perhaps a more serious weakness of this theory is that we cannot expect genes of small effect to be segregating at frequencies close to equilibrium if the original populations are of finite size. Hill and Robertson (1968) have studied the distribution of frequency of overdominant genes in finite lines selected on pure performance. The mean frequency in lines still segregating is not the equilibrium frequency, unless $\bar{q} = 0.5$, but is generally intermediate between $\bar{q}$ and 0.5, being closer to 0.5 if gene effects are small.

Fig. 5 shows clearly that when $Ni\sigma_d/\sigma_f$ is greater than one-half or so, a greater proportion of the possible advance from the equilibrium state is made if $\bar{q}$ has intermediate values. If $\bar{q}$ has extreme values, there is a high probability that the favourable allele, initially at higher frequency, will be fixed in both populations by chance.

Although approximately the same total advance is made with RRS and RST from initial equilibrium, the initial rate of advance and thus the overall pattern of change must be greatly different. In Fig. 6 the mean of the cross in succeeding generations is compared for RRS and RST

with $\bar{q}=0.5$ and three values of $Ns(=4Ni\sigma_d/\sigma_f$ for $\bar{q}=0.5)$. This figure illustrates the contrast in response rate. If Ns is small, it is shown in the appendix (setting $r=s$, $m=n$, and $p_0=q_0=\bar{q}$ in Eq. (A4)) that with initial equilibrium and RRS

$$\begin{aligned}\mu_t-\mu_0 &= NsaF_t^2\bar{q}^2(1-\bar{q})^2 \text{ approximately}\\ &= F_t^2\cdot Ni\sigma_d^2/\sigma_f\end{aligned} \tag{29}$$

where $F_t=1-(1-1/2N)^t$ or $e^{-t/2N}$, approximately, or the inbreeding coefficient estimated from pedigrees. Similarly, with RST, which is effective selection for additive genes, it can be shown with the same assumptions that

$$\mu_t-\mu_0=F_t\cdot Ni\sigma_d^2/\sigma_f \quad \text{approximately.} \tag{30}$$

Thus the selection advance is proportional to F^2 with RRS and F with RST. Also the half-life of the process, the time taken to get halfway to the limit (Robertson, 1960) will be $t=2.5N$ cycles or generations approximately, with RRS when $F^2=0.5$, and $t=1.4N$ generations, approximately, with RST when $F=0.5$. In the example in Fig. 6 with $Ns=1$

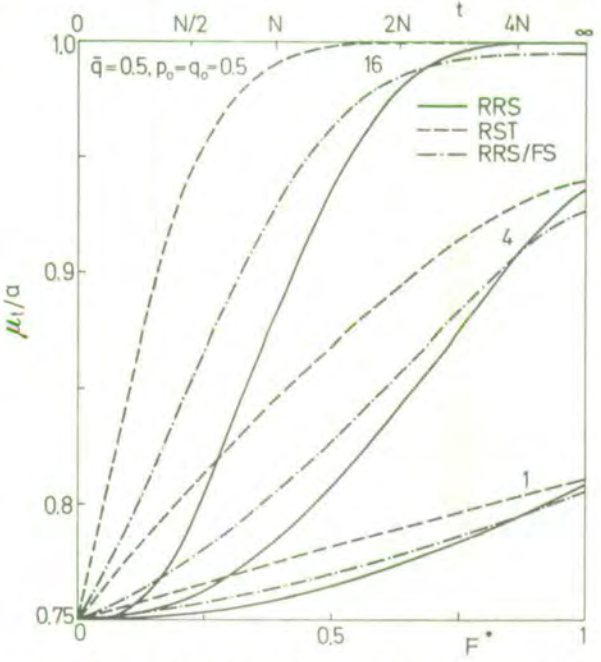

Fig. 6. Progress in the crossbreds from selection with overdominance and $\bar{q}=0.5$ for three values of Ns and RST, RRS with no prior inbreeding and RRS after one generation of full sibbing. Time is expressed as $F^*=1-e^{-t/2N}$ and the crossbred mean as a proportion of the effect a

these results hold fairly well. With RST the half-way point is passed between generations 10 and 11, corresponding to $1.25N$ and $1.375N$ generations respectively; and with RRS the corresponding generations are 19 and 20, or $2.375N$ and $2.5N$. With large Ns values the half-lives are seen to be reduced with both RRS and RST but are always shorter with RST. Again, assuming small gene effects, the response in a single generation is shown in the appendix for RRS and initial equilibrium to be

$$\mu_{t+1} - \mu_t = F_t(1 - F_t)\, i\sigma_d^2/\sigma_f, \quad \text{approximately.} \tag{31}$$

Similarly for RST, or additive selection

$$\mu_{t+1} - \mu_t = (1 - F_t)\, \tfrac{1}{2}\, i\sigma_d^2/\sigma_f, \quad \text{approximately.}$$

Thus, under these assumptions, the greatest rate of advance is made with RRS when $F_t = 0.5$. Then, incidentally, both RRS and RST have the same predicted rate of advance, but response is fastest with RST at the outset.

As Arthur (1964) has shown, an initial period of inbreeding enables response to be made immediately with RRS from an equilibrium position. It can be shown that if populations X and Y are inbred up to inbreeding coefficients F_0 and G_0, respectively, prior to selection, then the initial rate of advance with equilibrium becomes

$$\mu_1 - \mu_0 = \tfrac{1}{2} a[rF_0(1 - G_0) + sG_0(1 - F_0)]\, \bar{q}^2(1 - \bar{q})^2 \tag{32}$$

or if $F_0 = G_0$, and $r = s$,

$$\mu_1 - \mu_0 = F_0(1 - F_0)\, i\sigma_d^2/\sigma_f \tag{33}$$

which is identical to Eq. (31) but since no prior selection is involved, Eq. (33) does *not* require the assumption of small gene effects. A formal proof of Eq. (32) and (33) can be made by similar methods to those used in the appendix; however, the equations can be derived fairly easily in an intuitive manner. We have shown that

$$\delta\mu = \frac{a}{2}\,[rp(1 - p)\,(q - \bar{q})^2 + sq(1 - q)\,(p - \bar{q})^2]\,.$$

The average value of $p(1 - p)$ after inbreeding to level G_0 from an initial frequency of $\bar{q}$ is $(1 - G_0)\,\bar{q}(1 - \bar{q})$, for this is the within-line variance of gene frequency. Similarly, the average value of $(q - \bar{q})^2$ is $F_0\bar{q}(1 - \bar{q})$ for, with initial frequency $\bar{q}$, the quantity $(q - \bar{q})^2$ is the between-line variance. The genetic drift occurs independently in the two populations, so that the average value of $p(1 - p)\,(q - \bar{q})^2$ is $(1 - G_0)\,F_0\,\bar{q}^2(1 - \bar{q})^2$ and Eq. (32) follows immediately.

When gene effects are small, the total advance after inbreeding to level F_0 in each population becomes

$$\mu_L - \mu_0 = (1 - F_0^2)\, N i \sigma_d^2 / \sigma_f .$$

For example, after one generation of full sibbing in each population prior to selection, $F_0 = 1/4$. Therefore the early advance is $3/16\, i\sigma_d^2/\sigma_f$, or 3/4 of the maximum rate at $F = 1/2$, yet the total gain is reduced by only about 1/16. Results from RRS after a single generation of full sibbing are also included in Fig. 6, so that they can be compared with RST and RRS without prior inbreeding.

So far we have compared RRS and RST with the same values of Ns, yet RST requires only one segregating population and not two. Now it is conceivable, as has been mentioned earlier in the section on complete dominance, that facilities may limit the total number of individuals, $N + M$, which can be maintained. How then should our facilities be utilized? A solution is readily obtained for the model of small gene effects and initial equilibrium. Letting $r = s = ia/\sigma_f$, we have from Eq. (27)

$$\mu_L - \mu_0 = \left(M + N - \frac{2MN}{M + N - 1} \right) i\sigma_d^2/\sigma_f . \tag{34}$$

For $M + N$ constant, Eq. (34) has a relative minimum at $M \neq N$ and is maximized at $M = 0$ or $N = 0$. Thus the RST method is most efficient and the RRS method least efficient under these assumptions, differing by a factor of 2 in predicted advance.

A weakness in our analysis has been the assumption of instantaneous fixation with RST. Again using the model of small effects and initial equilibrium we test this approximation. We assume no selection in the tester and set $r = 0$ and $s = ia/\sigma_f$ in Eq. (27) and obtain

$$\mu_L - \mu_0 = \left(1 - \frac{2M}{2M + 2N - 1} \right) N i \sigma_d^2 / \sigma_f \tag{35}$$

where M is the size of the tester. If N is large relative to M, Eq. (35) becomes approximately

$$\mu_L - \mu_0 = (1 - M/N)\, N i \sigma_d^2 / \sigma_f . \tag{36}$$

For example, with $N = 16$ and $M = 2$ about 1/8 of the advance possible with instantaneous fixation and RST may not be realized.

Initial Disequilibrium. With overdominance and initial departure from the equilibrium frequency there are so many combinations of parameters that it is difficult to generalize. However in Figs. 7, 8, and 9 some typical results are presented, in which the selection limits predicted from RST and RRS are compared for the intermediate equilibrium frequency, $\bar{q}=0.5$, with one population at equilibrium (Fig. 7) and neither at equilibrium (Fig. 8), and for a more extreme equilibrium frequency, $\bar{q}=0.25$ (Fig. 9). In each graph results are given for two values of Ns.

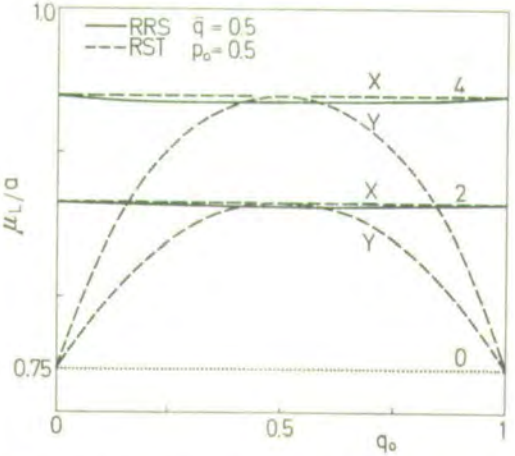

Fig. 7. The selection limit in the crossbreds, expressed as a proportion of the effect, a, with overdominance ($\bar{q}=0.5$) and two values of Ns. The initial frequency in X is $p_0=0.5$ and the initial frequency in Y is q_0. The population in which selection is practised with RST is also shown

From Fig. 7 we find that if $\bar{q}=0.5$ RST is more efficient if selection is practised in the population nearest the equilibrium frequency. This rule for $\bar{q}=0.5$ seems to hold quite generally, but has not yielded to algebraic proof. Also, we see in Figs. 7 and 8 that RST is never appreciably superior to RRS when $\bar{q}=0.5$ and comparisons are made at the same Ns value, and a similar result is observed in Fig. 9 when $\bar{q}\neq 0.5$. However no general rule has been found for identifying which population should be selected in an RST programme. As with partial dominance this depends not only on p_0, q_0, and $\bar{q}$ but also on Ns.

When there is initial disequilibrium, there is less difference in the rate of initial advance with RRS and RST programmes, for additive variance is immediately available with RRS. With $\bar{q}=0.5$ and $Ns=4$, some examples are given in Fig. 10. The rate of initial advance is always greater with RST (unless the population in which selection is practised has a very extreme frequency), but some immediate gains are made with

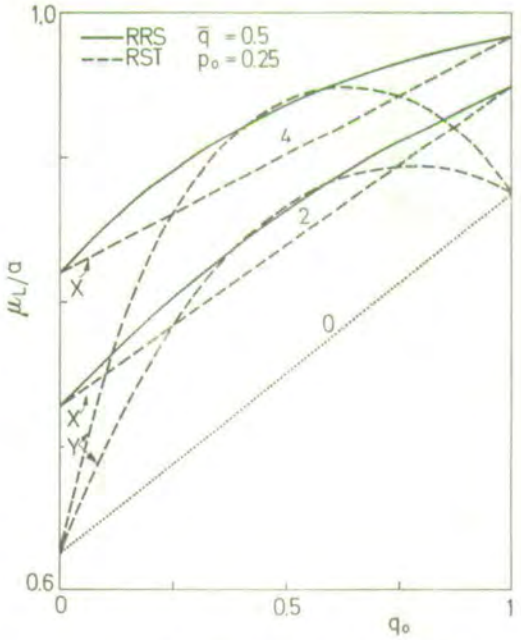

Fig. 8. As Fig. 7 but $\bar{q} = 0.5$, $p_0 = 0.25$

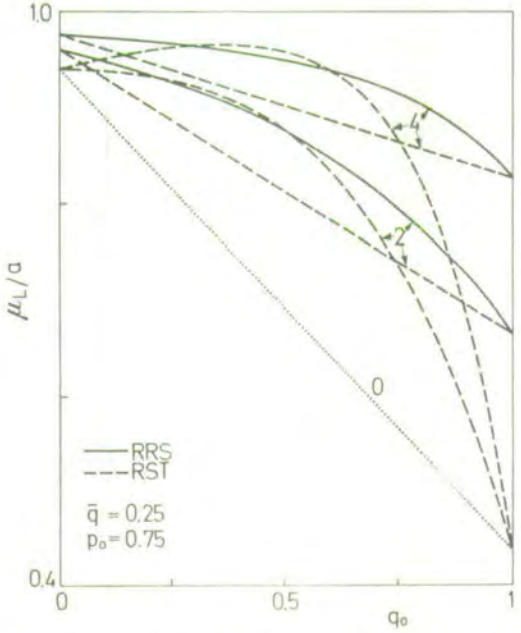

Fig. 9. As Fig. 7 but $\bar{q} = 0.25$, $p_0 = 0.75$

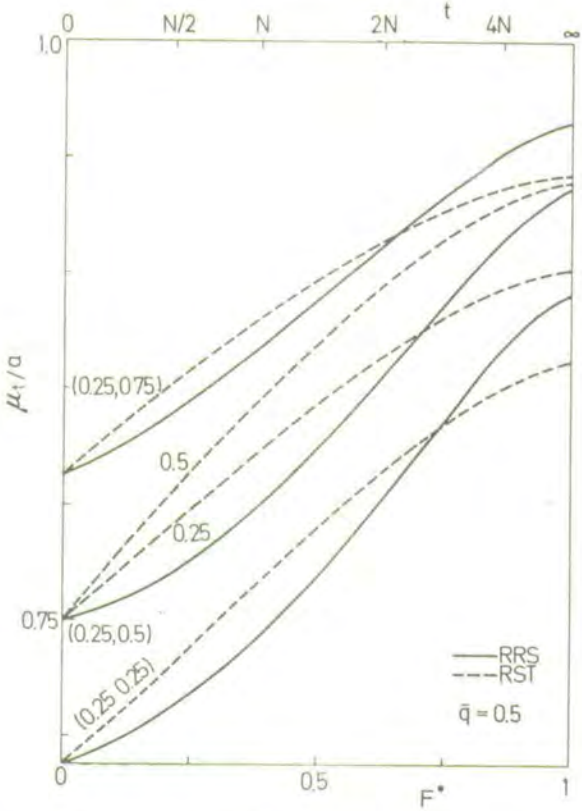

Fig. 10. Progress in the crossbreds from selection with overdominance and $\bar{q}=0.5$ and $Ns=4$ for three pairs of initial frequencies. The initial frequency of the selected population with RST is shown. Time is expressed as $F^*=1-e^{-t/2N}$ and the crossbred means as a proportion of the effect a

RRS. It is possible to analyze these situations in some detail for the model of small gene effects in the same manner as we have studied complete dominance and overdominance with initial equilibrium, but this will not be undertaken.

Discussion

Although the theory which has been developed is both very approximate and formally restricted to single genes, it is hoped that it gives some information on the problems of selection with subsequent line crossing. The main advance in the theory is, of course, the introduction of finite population size so that the selection limit defined is the expected

limit, not the maximum limit possible with fixation of only favourable combinations. Also by introducing finite population size we can draw a contrast between the initial rate of gain and the final limit. In fact the study of initial equilibrium with overdominance requires finite population theory.

A disappointing, but by no means surprising aspect of the results is that it is not possible to reach very general conclusions. We have considered each model of gene action in turn and find, in common with Comstock *et al.* (1949), Dickerson (1952), and Crow (1953), that the optimum breeding system is not the same for each model. For example with partial dominance RST is never very efficient, whereas with overdominance it may be the optimum system both for short term and long term gains. Of course, if epistasis were included in the study, further complications would inevitably be found. In addition some of the conclusions depend upon the size of gene effects on the quantitative trait under selection, and such information is almost completely lacking. Fortunately it appears from the results that the *relative* efficiency of the different methods discussed is not greatly influenced by the magnitude of effects, so in this respect our results may be of as much utility as those from infinite population studies.

There would be little benefit in entering a debate here about the nature of gene action found in practise – such speculations can be left to the breeder when setting up a programme. Unfortunately the experimental data available to him is unlikely to give unequivocal pointers to the genetics of the economic traits in which he is interested. One might argue that more knowledge of gene effects and equilibrium frequencies is necessary before a theory such as in this paper has any merit. However most breeders are used to designing programmes with insufficient information about the parameters, so new theory might still be considered beneficial. Since the relative efficiencies of the breeding systems depend so much on the nature of gene action, it is perhaps not surprising that experiments on different species or strains in the same species have not given clear-cut evidence about the utility of RRS, for example.

No attempt has been made to include in the theory refined variations on mating systems, particulary those possible with plants. The typical scheme adopted has been selection on the basis of progeny test performance, with random mating of individuals both within the strain producing the next generation and between the strains for test crossing. Incorporation of any other system, at least with random mating, should be possible in terms of the effective population size and selection coefficient per cycle, and the method has been outlined earlier in the paper. There are other, quite separate breeding systems, which merit further

analysis, perhaps within the framework of this paper. In particular what are the relative efficiencies of programmes based partly or largely on between-line selection? Arthur (1964) studied a model with recurrent cycles of inbreeding and between-line selection, but the gene effects were sufficiently large that the optimum limit was eventually reached.

The analysis has been basically in terms of single genes. However, without tight linkage it should be possible to extend them into a polygenic situation, merely by summation of variances and responses over loci. In order to simplify the analysis, we assumed that selective values remained constant from the early to the late generations, and since $s = ia/\sigma_f$ we are thereby assuming that σ_f also remains constant. As σ_f^2 contains genetic variance, this expectation is unlikely to be realized for both inbreeding and selection will modify the variance so that our results are biased. Qureshi and Kempthorne (1968) and Robertson (1970) use Monte Carlo methods to compute limits in single populations with many loci. They include the case of free recombination and find deviations from single locus theory, depending both on initial frequencies and gene effects. More important in our study is how the relative efficiencies of alternative systems are affected – a similar bias in each system is not important. The approximations may not be too serious as chance fixation of many favourable genes, especially those at extreme frequency, occurs in the early generations of selection, before the variance has changed too much. Inclusion of epistasis or linkage into the theory is in practise quite simple, if laborious, with Monte Carlo methods, but of course the number of possible parameter combinations increases enormously.

Summary

A theoretical comparison is made of alternative breeding systems which utilize only selection within lines to improve a cross between two strains. The schemes considered are selection on pure line performance (PLS) and selection on cross performance either by recurrent selection to an inbred tester (RST) or by reciprocal recurrent selection (RRS). This theory extends earlier comparisons in that the selected lines are assumed to be of finite size and predictions are made of the expected limit rather than the limit possible with fixation of only favourable combinations. A simple model of a single locus with two alleles with specified degree of dominance is used.

The selection limit is defined in terms of the combined parameter Ns, where N is the effective population size for a cycle of progeny testing and selection and $s = i\,a/\sigma_f$ where i is the standardized selection differential,

a the effect of the gene on the quantitative trait, and σ_f the standard deviation of progeny test means. Appropriate values of Ns can be calculated for other schemes such as mass selection and thus the results are quite general. For PLS and RST and for RRS with complete dominance, the limit could be predicted from diffusion equation approximations, and for all methods with small gene effects by simpler approximate methods. Other results were obtained using transition probability matrix iteration.

The optimum selection programme depends on the model of gene action. Comparisons made at the same value of Ns in each selected line gave the following results:

1. With *complete dominance* RRS is more effective than PLS and the efficiency of RST depends on the initial gene frequencies in the two strains. A higher limit with RRS is reached if selection is practised in the population with the higher initial frequency and the other is inbred. Then RST may be almost as efficient as RRS. Although there are some differences in rate of advance between the methods, the ranking does not change appreciably during selection.

2. With *partial dominance* RRS and PLS have similar efficiency, but RST is less efficient, particularly at high Ns values, since the optimum is attained only if both populations are fixed for the favourable allele.

3. With *overdominance and initial equilibrium* in each population PLS is not useful. RRS and RST give essentially the same limit for all values of the equilibrium frequency, but the initial rate of progress is much faster with RST. For genes of small effect the total advance up to cycle t is proportional to F_t with RST and F_t^2 with RRS, and the rates of response are proportional to $1 - F_t$ and $F_t(1 - F_t)$ respectively, where F_t is the inbreeding coefficient from the start of the programme. An initial period of inbreeding up to level F_0 in each line before commencing RRS gives an initial rate of advance proportional to $F_0(1 - F_0)$ and reduces the limit to $1 - F_0^2$, if effects are small.

4. With *overdominance and initial disequilibrium* the relative efficiency of RRS and RST depends on the initial frequencies, and general results are difficult to obtain. There is less difference in the rate of advance between the methods than with initial equilibrium.

Usually it will be possible to attain higher Ns values with PLS than RRS or RST, by using individual selection with reduced generation interval. Also, since only one non-inbred population has to be maintained with RST, it is possible that a larger Ns value can be utilized than with RRS. The rankings of the methods are likely to be affected.

Appendix — Response to Reciprocal Recurrent Selection when Selective Values are Small

From Eq. (8) we have the transition probability matrix $\boldsymbol{D}$, with elements

$$\begin{aligned} d_{(h,i,j,k)} = & \binom{m}{j}\left[\frac{h}{m} - \frac{r}{2}\,\frac{h}{m}\left(1-\frac{h}{m}\right)\left(\frac{i}{n}-\bar{q}\right)\right]^j \\ & \times\left[1-\frac{h}{m}+\frac{r}{2}\,\frac{h}{m}\left(1-\frac{h}{m}\right)\left(\frac{i}{n}-\bar{q}\right)\right]^{m-j} \\ & \times\binom{n}{k}\left[\frac{i}{n} - \frac{s}{2}\,\frac{i}{n}\left(1-\frac{i}{n}\right)\left(\frac{h}{m}-\bar{q}\right)\right]^k \\ & \times\left[1-\frac{i}{n}+\frac{s}{2}\,\frac{i}{n}\left(1-\frac{i}{n}\right)\left(\frac{h}{m}-\bar{q}\right)\right]^{n-k} \end{aligned} \tag{A1}$$

where $0 \leqq h, j \leqq m$; $0 \leqq i, k \leqq n$, and $m = 2M$ and $n = 2N$ in Eq. (8). Thus h and j are the numbers of A_1 alleles in population X and i and k the numbers in population Y at generations t and $t+1$ respectively so that $\boldsymbol{D}$ is square of dimension $(m+1)(n+1)$. A row of the matrix is identified by h and i, a column by j and k. Expanding $\boldsymbol{D}$ into terms up to order r or s, and rearranging, we obtain

$$\begin{aligned} d_{(h,i,j,k)} = & \binom{m}{j}\left(\frac{h}{m}\right)^j\left(1-\frac{h}{m}\right)^{m-j}\binom{n}{k}\left(\frac{i}{n}\right)^k\left(1-\frac{i}{n}\right)^{n-k} \\ & \times\left[1+\frac{r}{2}(h-j)\left(\frac{i}{n}-\bar{q}\right)+\frac{s}{2}(i-k)\left(\frac{h}{m}-\bar{q}\right)\right] \\ & + 0(r^2)+0(rs)+0(s^2). \end{aligned}$$

Let us define matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ with elements

$$a_{(h,i,j,k)} = \binom{m}{j}\left(\frac{h}{m}\right)^j\left(1-\frac{h}{m}\right)^{m-j}\binom{a}{k}\left(\frac{i}{n}\right)^k\left(1-\frac{i}{n}\right)^{n-k},$$

$$b_{(h,i,j,k)} = a_{(h,i,j,k)}(h-j)\left(\frac{i}{n}-\bar{q}\right),$$

$$c_{(h,i,j,k)} = a_{(h,i,j,k)}(i-k)\left(\frac{h}{m}-\bar{q}\right),$$

so that $\boldsymbol{D} = \boldsymbol{A} + \frac{r}{2}\boldsymbol{B} + \frac{s}{2}\boldsymbol{C} + 0(r^2) + 0(rs) + 0(s^2)$.

Let us also define vectors of $(m+1)(n+1)$ rows as follows:

$$V_1 : v_{1(h,i)} = \frac{hi}{mn} = pq\,,$$

$$V_2 : v_{2(h,i)} = \frac{h(m-h)\,i}{m^2 n}\left(\frac{i}{n} - \bar{q}\right) = p(1-p)\,q(q-\bar{q})\,,$$

$$V_3 : v_{3(h,i)} = \frac{i(n-i)\,h}{mn^2}\left(\frac{h}{m} - \bar{q}\right) = q(1-q)\,p(p-\bar{q})\,,$$

$$V_4 : v_{4(h,i)} = \frac{h(m-h)\,i(n-i)}{m^2 n^2} = p(1-p)\,q(1-q)$$

where $p = h/m$, $q = i/n$. It can be shown that

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{V}_1 &= \boldsymbol{V}_1\,,\\
\boldsymbol{B}\boldsymbol{V}_1 &= -\boldsymbol{V}_2\,,\\
\boldsymbol{C}\boldsymbol{V}_1 &= -\boldsymbol{V}_3\,,\\
\boldsymbol{A}\boldsymbol{V}_2 &= \frac{m-1}{m}\boldsymbol{V}_2 + \frac{m-1}{mn}\boldsymbol{V}_4\,,\\
\boldsymbol{A}\boldsymbol{V}_3 &= \frac{n-1}{n}\boldsymbol{V}_3 + \frac{n-1}{mn}\boldsymbol{V}_4\,,\\
\boldsymbol{A}\boldsymbol{V}_4 &= \frac{(m-1)(n-1)}{mn}\boldsymbol{V}_4\,.
\end{aligned}$$

The probability that a cross between X and Y at generation t has progeny A_1A_1 is given by elements of the vector $\boldsymbol{V}_1^{(t)} = \boldsymbol{D}^t\boldsymbol{V}_1$ and the joint probability of fixation of A_1 in both lines by $\boldsymbol{V}_1^{(\infty)} = \lim_{t\to\infty}\boldsymbol{D}^t\boldsymbol{V}_1$. We now use the above relationships to derive these quantities, but ignoring high order terms in r and s. We assume in the following equations that r and s are of similar order of magnitude.

$$\boldsymbol{V}_1^{(1)} = (\boldsymbol{A} + r\boldsymbol{B} + s\boldsymbol{C})\,\boldsymbol{V}_1 = \boldsymbol{V}_1 - \frac{r}{2}\boldsymbol{V}_2 - \frac{s}{2}\boldsymbol{V}_3 + 0(s^2)\,,$$

$$\begin{aligned}
\boldsymbol{V}_1^{(2)} = (\boldsymbol{A} + r\boldsymbol{B} + s\boldsymbol{C})\,\boldsymbol{V}^{(1)} &= \boldsymbol{V}_1 - \frac{r}{2}\left(1 + \frac{m-1}{m}\right)\boldsymbol{V}_2 - \frac{s}{2}\left(1 + \frac{n-1}{n}\right)\boldsymbol{V}_3\\
&\quad - \frac{r}{2}\left(\frac{m-1}{mn}\right)\boldsymbol{V}_4 - \frac{s}{2}\left(\frac{n-1}{mn}\right)\boldsymbol{V}_4 + 0(s^2)\,,
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{V}_1^{(3)} = \boldsymbol{V}_1 &- \frac{r}{2}\left[1 + \frac{m-1}{m} + \left(\frac{m-1}{m}\right)^2\right]\boldsymbol{V}_2 - \frac{s}{2}\left[1 + \frac{n-1}{n} + \left(\frac{n-1}{n}\right)^2\right]\boldsymbol{V}_3\\
&- \frac{r}{2}\left[\frac{m-1}{mn} + \frac{(m-1)^2}{m^2 n} + \frac{(m-1)^2(n-1)}{m^2 n^2}\right]\boldsymbol{V}_4\\
&- \frac{s}{2}\left[\frac{n-1}{mn} + \frac{(n-1)^2}{mn^2} + \frac{(m-1)(n-1)^2}{m^2 n^2}\right]\boldsymbol{V}_4 + 0(s^2)
\end{aligned}$$

and so on. In general, if we write

$$V_1^{(t)} = V_1 - \tfrac{1}{2}\left[\alpha_t r V_2 + \beta_t s V_3 + \gamma_t r V_4 + \delta_t s V_4\right] + 0(s^2), \qquad \text{(A 2)}$$

we obtain, for example, the following recurrence relationships

$$\gamma_{t+1} = \left(\frac{m-1}{mn}\right)\alpha_t + \frac{(m-1)(n-1)}{mn}\gamma_t$$

with initial conditions $\alpha_0 = \beta_0 = \gamma_0 = \delta_0 = 0$.

These have solutions and limiting values

$$\alpha_t = m\left[1 - \left(\frac{m-1}{m}\right)^t\right], \quad \lim_{t\to\infty}\alpha_t = m,$$

$$\beta_t = n\left[1 - \left(\frac{n-1}{n}\right)^t\right], \quad \lim_{t\to\infty}\beta_t = n,$$

$$\gamma_t = \frac{m(m-1)\left[1-\left(\frac{m-1}{m}\right)^t\right] - mn\left(\frac{m-1}{m}\right)^t\left[1-\left(\frac{n-1}{n}\right)^t\right]}{m+n-1},$$

$$\lim_{t\to\infty}\gamma_t = \frac{m(m-1)}{m+n-1},$$

$$\delta_t = \frac{n(n-1)\left[1-\left(\frac{n-1}{n}\right)^t\right] - mn\left(\frac{n-1}{n}\right)^t\left[1-\left(\frac{m-1}{m}\right)^t\right]}{m+n-1}$$

$$\lim_{t\to\infty}\delta_t = \frac{n(n-1)}{m+n-1}.$$

The vector of probability of fixation of A_1 in both lines is obtained by substitution in Eq. (A2) and is

$$V_1^{(\infty)} = V_1 - \frac{1}{2}\left\{mrV_2 + nsV_3 + \left[mr + ns - \frac{mn(r+s)}{m+n-1}\right]V_4\right\} + 0(s^2).$$

In terms of gene frequencies, the joint chance of fixation $w(p_0, q_0)$ is then

$$\begin{aligned} w(p_0, q_0) = p_0 q_0 &- m\frac{r}{2}p_0(1-p_0)q_0(q_0-\bar{q}) - n\frac{s}{2}q_0(1-q_0)p_0(p_0-\bar{q}) \\ &- \frac{1}{2}\left[mr + ns - \frac{mn(r+s)}{m+n-1}\right]p_0(1-p_0)q_0(1-q_0) + 0(s^2). \end{aligned} \qquad \text{(A 3)}$$

The other joint probabilities of fixation, such as A_2 in each population $x(1-p_0, 1-q_0)$, are obtained using Eq. (A 3) and the marginal prob-

abilities of fixation, which can be shown to be

$$u(p_0) = p_0 - m\frac{r}{2} p_0(1-p_0)(q_0-\bar{q}) + 0(s^2),$$

$$v(q_0) = q_0 - n\frac{s}{2} q_0(1-q_0)(p_0-\bar{q}) + 0(s^2).$$

The selection limit is given by $\mu_L = a[1-(1-\bar{q})\, w(p_0, q_0) - \bar{q}x(1-p_0, 1-q_0)]$ and, after substituting, the total advance becomes

$$\mu_L - \mu_0 = \frac{1}{2} a \left\{ mr p_0(1-p_0)(q_0-\bar{q})^2 + ns q_0(1-q_0)(p_0-\bar{q})^2 + \left[mr + ns - \frac{mn(r+s)}{m+n-1} \right] p_0(1-p_0)\, q_0(1-q_0) \right\} + 0(s^2).$$

In order to specify the response at intermediate generations it is convenient to let $F_t = 1 - \left(\frac{m-1}{m}\right)^t$, $G_t = 1 - \left(\frac{n-1}{n}\right)^t$ so that for example, $\alpha_t = mF_t$, $\gamma_t = \dfrac{m[(m-1)F_t - n(1-F_t)G_t]}{m+n-1}$ and

$$\mu_t - \mu_0 = \frac{1}{2} a \left\{ mrF_t p_0(1-p_0)(q_0-\bar{q})^2 + nsG_t q_0(1-q_0)(p_0-\bar{q})^2 + \frac{mr[(m-1)F_t - n(1-F_t)G_t] + ns[(n-1)G_t - m(1-G_t)F_t]}{m+n-1} \times p_0(1-p_0)\, q_0(1-q_0) \right\} + 0(s^2). \tag{A4}$$

If $m = n$ and $r = s$ and n is sufficiently large that $m+n-1 \sim m+n$,

$$\mu_t - \mu_0 = \tfrac{1}{2} ansF_t [p_0(1-p_0)(q_0-\bar{q})^2 + q_0(1-q_0)(p_0-\bar{q})^2 + F_t p_0(1-p_0)\, q_0(1-q_0)] + 0(s^2) \tag{A5}$$

and

$$\mu_{t+1} - \mu_t = \tfrac{1}{2} as(1-F_t)[p_0(1-p_0)(q_0-\bar{q})^2 + q_0(1-q_0)(p_0-\bar{q})^2 + 2F_t p_0(1-p_0)\, q_0(1-q_0)] + 0(s^2). \tag{A6}$$

Proof of Convergence. We have not shown that $w(p_0, p_0) - p_0 q_0$ of Eq. (A2) actually converges to the quantity found as r and s tend to zero. We note that $\boldsymbol{D}$ has elements which are polynomials in r and s, hence then are all $\boldsymbol{D}^t$, including $\lim_{t\to\infty} \boldsymbol{D}^t$. But the elements of $\lim_{t\to\infty} \boldsymbol{D}^t \boldsymbol{V}_1 = \boldsymbol{V}_1^{(\infty)}$ are bounded in the range $0 \leqq v_{1(h,i)}^{(\infty)} \leqq 1$, since $\boldsymbol{V}_1^{(\infty)}$ is a vector of probabilities. Thus we can expand $\boldsymbol{V}_1^{(\infty)}$ in a series of vectors

$$\boldsymbol{V}_1 = \boldsymbol{E}_{00} + r\boldsymbol{E}_{10} + s\boldsymbol{E}_{01} + r^2\boldsymbol{E}_{20} + rs\boldsymbol{E}_{11} + s^2\boldsymbol{E}_{02} + \cdots$$

for all $r, s, -\infty < r, s < \infty$, where $\boldsymbol{E}_{ij}$ are vectors of coefficients. Therefore $-\infty < \boldsymbol{E}_{ij} < \infty$ for all i, j, and if r and s are of the same order, say $r = \sigma s$, $0 < \sigma < \infty$

$$\lim_{s \to 0} \left(\frac{\boldsymbol{V}_1^{(\infty)} - \boldsymbol{E}_{00}}{s} \right) = \sigma \boldsymbol{E}_{10} + \boldsymbol{E}_{01} + \lim_{s \to 0} [s(\sigma^2 \boldsymbol{E}_{20} + \sigma \boldsymbol{E}_{11} + \boldsymbol{E}_{02} + \cdots)]$$

$$= \sigma \boldsymbol{E}_{10} + \boldsymbol{E}_{01},$$

and we can ignore all terms of higher order as a first approximation.

References

Arthur, J. A.: Investigation of population structure with recurrent selection. Unpublished Ph. D. Thesis, University of California, Davis (1964).

Bell, A. E., C. H. Moore, and D. C. Warren: The evaluation of new methods for the improvement of quantitative characteristics. Cold Spr. Harb. Symp. Quant. Biol. **20**, 197–211 (1955).

Bowman, J. C.: Selection for heterosis. Anim. Breed. Abstr. **27**, 261–273 (1959).

Comstock, R. E., H. F. Robinson, and P. H. Harvey: A breeding procedure designed to make maximum use of both general and specific combining ability. Agron. J. **41**, 360–367 (1949).

Cress, C. E.: A comparison of recurrent selection schemes. Genetics **54**, 1371–1379 (1966).

— Reciprocal recurrent selection and modifications in simulated populations. Crop Sci. **7**, 561–567 (1967).

Crow, J. F.: Theoretical considerations of reciprocal recurrent selection and recurrent selection compared to other breeding methods. Proc. Amer. Poultry Breeders' Roundtable, 1953.

Dickerson, G. E.: Inbred lines for heterosis tests? *Heterosis*, pp. 330–351, ed J. W. Gowen. Ames: Iowa State College Press 1952.

Ewens, W. J.: Numerical results and diffusion approximations in a genetic process. Biometrika **50**, 241–249 (1963).

Falconer, D. S.: *Introduction to quantitative genetics*. Edinburgh-London: Oliver and Boyd 1960.

Griffing, B.: Theoretical consequences of truncation selection based on the individual phenotype. Aust. J. Biol. Sci. **13**, 307–343 (1960).

Haldane, J. B. S.: A mathematical theory of natural and artificial selection. VII. Selection intensity as a function of mortality rate. Proc. Camb. Phil. Soc. **27**, 131–136 (1931).

Hill, W. G.: On the theory of artifical selection in finite populations. Genet. Res. **13**, 143–163 (1969a).

— The rate of selection advance for non-additive loci. Genet. Res. **13**, 165–173 (1969b).

—, and A. Robertson: The effects of inbreeding at loci with heterozygote advantage. Genetics **60**, 615–628 (1968).

Hull, F. H.: Recurrent selection for specific combining ability in corn. J. Amer. Soc. Agron. **37**, 134–145 (1945).

Kimura, M.: Some problems of stochastic processes in genetics. Ann. Math. Statist. **28**, 882–901 (1957).

— Diffusion models in population genetics. J. Appl. Prob. **1**, 177–232 (1964).

Kojima, K.: Effects of dominance and size of population on response to mass selection. Genet. Res. **2**, 177–188 (1961).

—, and T. M. Kelleher: A comparison of purebred and crossbred selection schemes with two populations of *Drosophila subobscura*. Genetics **48**, 57–72 (1963).

Latter, B. D. H.: The response to artificial selection due to autosomal genes of large effect. I. Changes in gene frequency at an additive locus. Aust. J. Biol. Sci. **18**, 585–598 (1965).

Ohta, T.: Effect of initial linkage disequilibrium and epistasis on fixation probability in a small population, with two segregating loci. Theor. Appl. Genet. **38**, 243–248 (1968).

Qureshi, A. W., and O. Kempthorne: On the fixation of genes of large effects due to continued truncation selection in small populations of polygenic systems with linkage. Theor. Appl. Genet. **38**, 249–255 (1968).

Rasmuson, M.: Reciprocal recurrent selection. Results of three model experiments on *Drosophila* for improvement of quantitative characters. Hereditas **42**, 397–414 (1956).

Robertson, A.: A theory of limits in artificial selection. Proc. Roy. Soc. Lond. B. **153**, 234–249 (1960).

— Selection for heterozygotes in small populations. Genetics **47**, 1291–1300 (1962).

— A theory of limits in artificial selection with many linked loci. (this volume) (1970).

6

Population structure in artificial selection programmes : simulation studies

by

Fernando E. Madalena and William G. Hill

Genet. Res., Camb. (1972), **20**, *pp.* 75–99
With 6 text-figures
Printed in Great Britain

Population structure in artificial selection programmes: simulation studies

By F. E. MADALENA† and W. G. HILL
Institute of Animal Genetics, Edinburgh EH9 3JN

(*Received 11 January 1972*)

SUMMARY

A simulation study was undertaken of methods of subdividing populations into several small sublines and utilizing the variances generated between lines by selecting among them. Crosses of chosen lines were made, and either selection was continued in a single large population (single cycle) or the population was subdivided again (repeated cycles). As a control for the efficiency of these schemes, a single large population was maintained and selected at the same intensity from the outset. Simple models were used of additive or completely dominant genes, usually of equal effect and equally spaced on a single chromosome.

The single and repeated cycle structures give similar results, but the repeated cycle structure is more extreme.

With additive models intense selection between lines gives short-term advances, but causes a reduction in the limit when compared with a single population. The effect on the limit is greatest with free recombination, very small with complete linkage. If no selection is practised between lines the limit is unaffected, but takes longer to attain.

With complete dominance, and the recessive allele initially at low frequency, greater responses from selection are obtained within sublines than in the large population, large gains are made from selection between sublines, and a higher limit can be reached. If the recessive allele is at high initial frequency the subdivision is not beneficial.

Some simple theory is developed to explain these results. It is concluded that subdivision and crossing schemes are unlikely to be very useful except for elimination of deleterious recessive genes.

1. INTRODUCTION

In the ideal selection programme rapid response would be made from the outset, and would continue until all the useful genetic variation in the source material had been incorporated. Unfortunately these objectives are partly incompatible since selected populations are necessarily of finite size. Rapid short-term gains can be made by selecting a very small proportion of the population for breeding the next generation, but many favourable genes will be lost by chance and the limit will be reduced. Dempster (1955) and Robertson (1960) showed theoretically that for single genes the limit is maximized when 50 % of the population are selected each

† Present address: Facultad de Agronomia, Universidad de la Republica, Paysandu, Uruguay.

generation. When linkage effects are important, rather more than 50% should be chosen (Hill & Robertson, 1966; Robertson, 1970*a*). More intense selection should be practised if the total advance is to be maximized in a specified, finite, number of generations (Robertson, 1970*b*), or if higher economic weight is given to early response (James, 1972). But in an attempt to avoid the conflict between short-term and long-term gains we should look at other breeding systems, such as structured or subdivided populations.

The structure of Mendelian populations has long been recognized as an important factor in evolution (Wright, 1951). Its effects on the progress from artificial selection have received less attention, except in breeding plans designed to exploit non-additive variation for improvement of line crosses. However Baker & Curnow (1969) considered populations divided into small sublines, and compared the rates of response and variance between lines for different sizes of the sublines and for alternative genetic models. They predicted that useful gains could be made even with small sublines, and then considerable further response could be obtained by selection between lines. Wright (1939) proposed a structure of repeated cycles of subdividing the population and practising within and between-line selection and crossing. He considered this method would be effective in preventing the loss by recombination of favourable epistatic combinations in cross-fertilizing species, and with a model of multiple 'peaks' of desirability in relation to gene frequencies, drift could allow the population as a whole to move to new peaks after crossing (Wright, 1951). Baker & Curnow (1969) did not investigate the effects of reselection from line crosses.

Some relevant theory is known however. With a model of independent additive genes Robertson (1960) showed that if m replicate lines were selected to fixation with size N each, crossed together and selected as a single population with size Nm, the same final limit would be attained as in a single population selected throughout at the same intensity with size Nm. Maruyama (1970) generalized these results for additive genes by showing that any subdivision of the total population gives the same selection limit, regardless of when crossing or migration occurs, so long as this happens without a change in mean gene frequency in the total population, i.e. without selection between lines. This generalization can also be derived from a formula given by Pollak (1966). Robertson's (1960) result for crosses of fixed lines holds approximately with dominance, but the subdivision structure gives a slightly higher limit when the recessive allele is favoured, a slightly lower limit when the dominant allele is favoured.

However, in structures in which the population is subdivided into lines of smaller size, the additive genetic variance within lines and consequently the response to selection are reduced by random drift. Thus unless selection between lines is practised the limit will take longer to reach in a subdivided population, except perhaps if the variability derives from low-frequency recessive genes when the additive variance may increase with initial inbreeding (Robertson, 1952). Since inbreeding increases variability between lines which can be utilized accurately by selection of the lines on mean performance it may be possible to design subdivided systems to obtain higher rates of advance and perhaps limits than by selection in a single population.

Experimental studies of gains from artificial selection in population structures involving between-line selection have been made by Bowman & Falconer (1961), Hill (1963), Madalena (1970) and Goodwill (1971). While the results obtained in these experiments with different traits of various species are not the same, in no case are large gains obtained from between-line selection and crossing, relative to selection in single populations.

In this paper a theoretical study has been made of structures utilizing between-line selection similar to those proposed by Wright (1939), and a preliminary report has already appeared (Hill & Madalena, 1969). Although we have not considered epistatic loci, linkage has been included, so that we can carry further the results of Robertson (1960) and Maruyama (1970). Monte Carlo simulation techniques have been used throughout; simple approximations using selective values at a single locus are not adequate, for the selective value at the locus during between-line selection is very much affected by segregation at the other loci.

In all comparisons which we make between selection schemes, the same total number of individuals (Q) are recorded each generation, either in one population with Q measured, or, say, 8 with $Q/8$ measured in each. Only in this way can a fair comparison between alternatives be made in terms of expense of measurement or utilization of facilities. However, we ignore biological difficulties, such as a decline in reproductive performance due to inbreeding.

2. METHODS

(i) *Design of population structures*

The structures studied are shown diagrammatically in Fig. 1. These are the *single-cycle structure* (Fig. 1*a*) in which one cycle of subdivision into small lines and intercrossing of selected lines is followed by selection thereafter in a single large population; and the *repeated-cycle structure* (Fig. 1*b*) in which a new set of lines are started from the intercross of the initial lines and the same procedure of inbreeding and crossing repeated.

In both systems the first cycle started at generation 0 with sampling of M individuals at random into each of m replicate lines from a base population in Hardy–Weinberg and linkage equilibrium. These M individuals were scored for a quantitative trait which was a function of their genotype and environmental error. The best N were chosen by truncation selection to be parents of the next generation and M progeny were bred.

Selection at this intensity (N/M) was continued for T generations. At generation T between-line selection was practised on the mean phenotype of the M individuals in the line, and the best v from the m lines chosen. In these v lines, within-line selection was again practised at the same intensity as before to give N individuals in each, a total of Nv, for crossing. These Nv individuals were randomly mated and selfed as if they were a single population to give a total of Q progeny. Thus both cross and 'pure' line progeny were formed, with the total number of chromosomes sampled from any line following a multinomial distribution. To allow recombination among

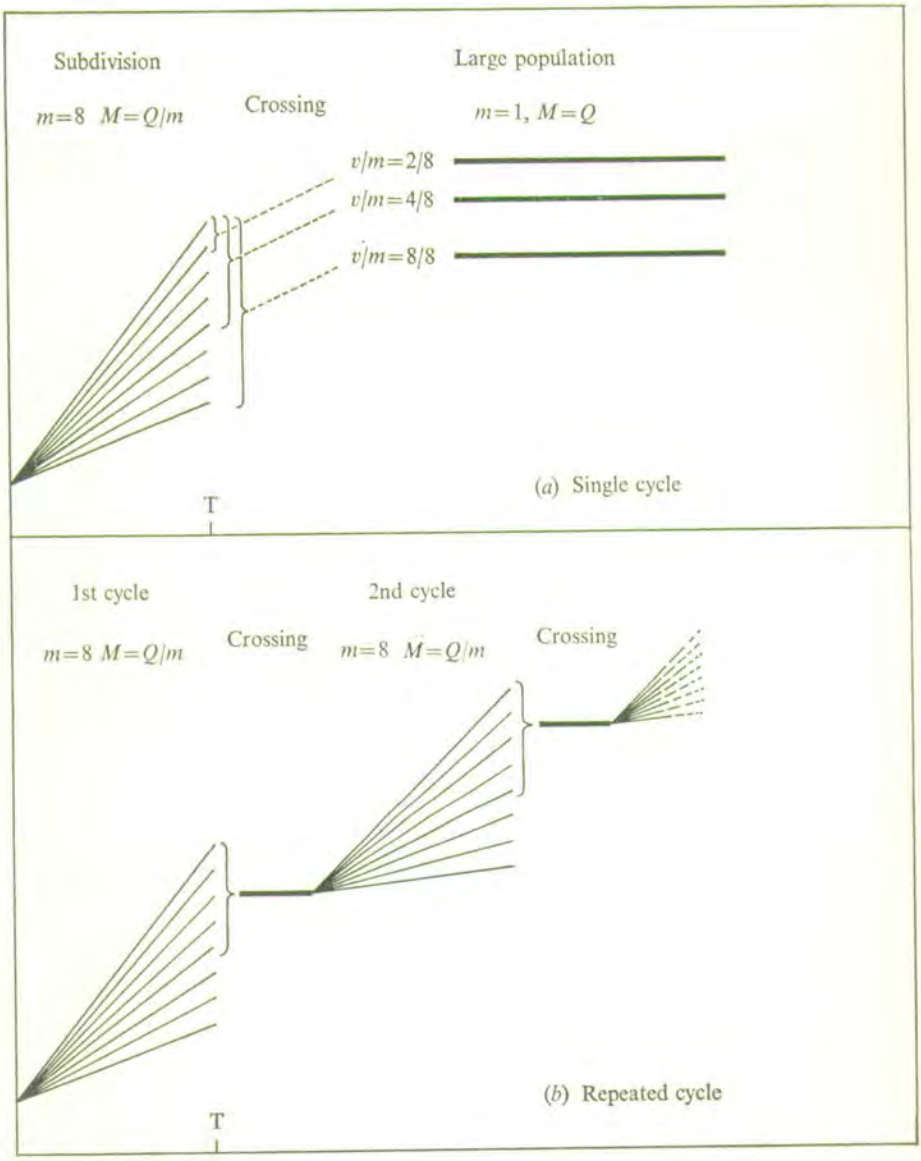

Fig. 1. The structures studied. (*a*) Single cycle: subdivision of the Q individuals measured into m lines of $M = Q/m$ individuals each, selection within lines of N individuals (a proportion N/M) for T generations. At generation T, selection between lines and crossing v selected lines to form a single population with Q individuals recorded and Nm selected (again a proportion N/M) until fixation. (*b*) Repeated cycle: repetition of cycles each of subdivision, selection within lines, selected between lines and crossing.

genes from different parent lines, these Q individuals were mated at random, without selection, and gave Q progeny at generation $T+2$.

A new cycle could therefore start at generation $T+2$. In the one-cycle structure, however, the cross population was maintained as a single large population of size Q and selected with intensity $Nm/Q(= N/M)$. In the repeated cycle structure the

Q individuals at generation $T+2$ were subdivided randomly into m lines, again of size M, and the process repeated. Thus each cycle (including the first) lasted $T+2$ generations, with T generations of within-line selection preceding the between-line selection, 1 generation of within-line selection in the chosen lines, and 1 generation without selection following crossing.

The symbols are summarized below:

Q = total number of individuals measured per generation ($Q = Mm$), and is the same for all systems;
m = number of replicate lines;
M = number of individuals measured per line;
N = number of parents selected in each line, so intensity of within line selection $= N/M$;
v = number of lines selected, so intensity of between-line selection $= v/m$;
T = number of generations of sublining before between-line selection.

A single large population (denoted L) in which mass selection was practised without subdivision was maintained as a control selection system. Each generation Nm individuals were chosen from a total of Q recorded, so that the L line had a size m times as large as the sublines, but had the same selection intensity as that used within lines. It was thus maintained in the same way as the large population after line crossing in the single cycle structure.

(ii) *Genetic model*

Individuals were assumed to be monecious diploids, in which random mating was accompanied by random selfing. The following parameters describe the genetic model:

n = number of loci affecting the character;
a = difference between the homozygotes at a locus in their effect on the character, with all loci having two alleles and additive or completely dominant genes, but no epistasis;
q = initial frequency of favourable allele;
c = recombination fraction between adjacent loci, with all loci equally spaced on a single chromosome;
σ^2 = variance of normally distributed environmental error.

For additive genes the initial heritability of the trait, h^2, is given by

$$h^2 = \tfrac{1}{2}na^2q(1-q)/[\tfrac{1}{2}na^2q(1-q)+\sigma^2].$$

In our runs we have typically taken $n = 5$, $a/\sigma = 0{\cdot}5$ so $h^2 = q(1-q)/[q(1-q)+1{\cdot}6]$. With an initial frequency of $q = 0{\cdot}2$, then $h^2 = 0{\cdot}1/1{\cdot}1 \sim 0{\cdot}1$, which changes during the course of selection, tending to increase initially due to selection but finally to decrease due to inbreeding. We have generally used heritabilities of this order; although they are low, they refer to single chromosomes.

In any generation chromosomes were paired in the order they were produced, to form genotypes. Their genotypic value was computed, an environmental deviation

added and truncation selection practised. The first chromosome for the next generation was obtained by choosing one of the selected parents at random, and performing a random walk (conceptually) along its chromosomes to permit recombination. This process was repeated until the required number of chromosomes were obtained. The whole experiment of sublining, selection, crossing, etc., was replicated 100 or so times for each set of parameters. In each replicate, lines were carried for 80 generations or until fixation, which usually occurred earlier, although limits are denoted '∞' in the tables.

Simulation was carried out on the Edinburgh Regional Computing Centre's KDF 9 computer. Inner loops in machine language were kindly written for us by Dr J. A. Burns.

3. RESULTS

In most of the genetic models which we have studied, where we have found a difference in rate of response or limit to selection between the large population and single cycle structure, we have also found a difference of the same direction, but not size, in rates or limits between the large population and repeated cycle structure. Most of our results therefore refer to the single cycle structure, since by using the same set of sublines to originate the subsequent large lines after different times and intensities of between-line selection, a greater range of parameters could be investigated with the single cycle than the repeated cycle structure for a given computing cost. For example, a set of $M = 8$ sublines was generally used to initiate 9 subsequent large lines, comprising three values of T (usually 1, 3 and 7), each with three values of v (usually 2, 4 and 8). In addition, a positive correlation is induced between the responses in the populations started from the same set of single cycle lines, so that the variance in response between them is reduced.

We shall investigate in turn those 'structural' parameters, such as the number of sublines, which can be controlled by the breeder. In each case we consider how the comparisons between alternative schemes are affected by the genetic model, which is outside the breeder's control. But since the results differ markedly for additive and non-additive models, we shall discuss these separately.

(i) *Single-cycle structure: additive model*

(*a*) *Between-line selection.* A typical result is shown in Fig. 2 for a simple model of five loci of equal effects and initial frequency 0·2 at each. The mean of the selected trait is then a linear function of the mean gene frequency, which is plotted. Prior to crossing, the figure shows the mean performance of all replicate sublines, which soon falls behind that of the large population as the within-line variance of the small lines is reduced. When all sublines are used at generation 3 to make the cross ($v/m = 8/8$) the mean advance lags behind that of the single population, and is furthest behind immediately following line crossing. However, the new synthetic population reaches about the same limit, within the range of sampling error. From Maruyama's (1970) theory we would expect this result for independent loci, but it seems to hold even for those which are tightly linked. Similarly, for other runs we

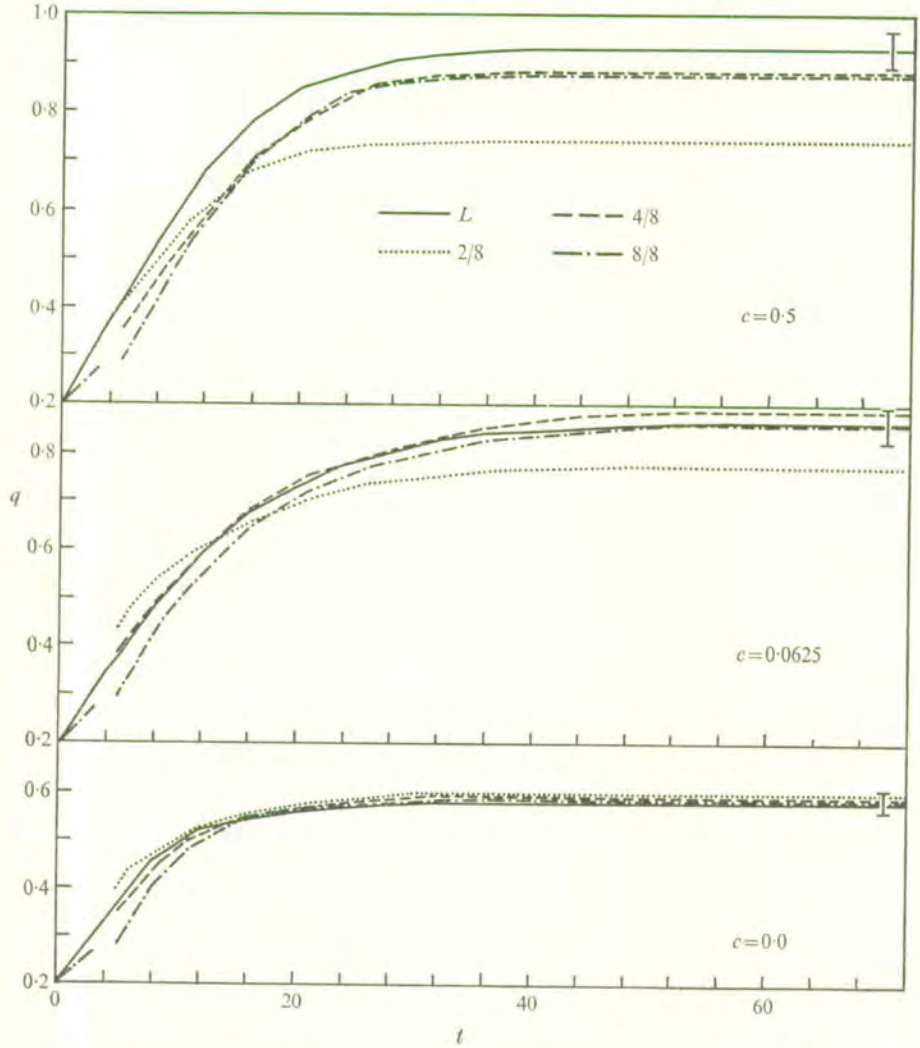

Fig. 2. Comparison between selection in a large population (L) and alternative intensities of between line selection (v/m = 2/8, 4/8 and 8/8) in the single cycle structure with $Q = 40$, $m = 8$, $M = 5$, $N = 2$ for an additive model with $n = 5$, $a/\sigma = 0{\cdot}5$, $q = 0{\cdot}2$ and recombination fraction, c. The mean gene frequencies are shown; these are for the mean of all sublines prior to between-line selection at generation 3 ($= T$) and the blank at generation 4 denotes the random mating following crossing. A range of length approximately 2 standard errors is shown for the difference between L and alternative structures at the limit.

have made for additive models, there is never an important difference between the limits obtained in the single population and in the two-cycle structure when there is no between-line selection. When selection is practised between lines we see (Fig. 2) that following crossing the mean of the cross may exceed that of the large population and remain ahead for a few generations. However, with intense between-line selection ($v/m = 2/8$) the limit for the single-cycle structure is lower than the limit for the single large population (L), except when the genes are very tightly linked, when there is no difference.

Table 1. *Effect of selection intensity between lines, initial gene frequency and linkage in a single-cycle structure with an additive model*

($Q = 40$, $m = 8$, $N/M = 2/5$, $T = 3$, relative response ± S.E. or mean frequency in $L(\bar{q}_L)$.)

$n = 5, a/\sigma = 0{\cdot}5$		$c = 0{\cdot}5$				$c = 0{\cdot}0625$				$c = 0{\cdot}0$			
q	v	$t = 5$	10	20	∞	$t = 5$	10	20	∞	$t = 5$	10	20	∞
0·1	2	6 ± 19	−6	−27	−35 ± 4	24 ± 14	−1	−15	−19 ± 5	−8 ± 6	−8	−18	−19 ± 8
	4	−18 ± 14	−6	−6	−11 ± 6	−12 ± 13	−12	−9	−5 ± 6	−22 ± 10	−13	−8	−11 ± 8
	8	−50 ± 10	−22	−4	−7 ± 5	−60 ± 9	−24	−14	−4 ± 5	−56 ± 9	−23	−7	−6 ± 8
	$\bar{q}_L$	0·21	0·36	0·60	0·73	0·24	0·37	0·56	0·68	0·23	0·34	0·41	0·42
0·5	2	−9 ± 8	−14	−6	−5 ± 3	8 ± 13	−10	−10	−7 ± 4	17 ± 10	4	−3	−3 ± 6
	4	−22 ± 7	−9	−3	0 ± 0	−7 ± 10	−9	1	1 ± 1	−5 ± 9	−4	−4	−3 ± 6
	8	−59 ± 7	−18	−5	0 ± 0	−44 ± 6	−19	−1	−1 ± 2	−50 ± 5	−20	−7	−5 ± 6
	$\bar{q}_L$	0·75	0·88	0·98	1·00	0·74	0·88	0·97	1·00	0·73	0·85	0·91	0·92
0·7	2	15 ± 11	5	2	0 ± 0	30 ± 11	2	−6	−3 ± 2	28 ± 9	2	1	−1 ± 3
	4	−16 ± 10	−6	1	0 ± 0	4 ± 11	−4	−6	0 ± 0	1 ± 9	0	−1	−1 ± 3
	8	−43 ± 7	−9	−2	0 ± 0	−34 ± 10	−9	−2	0 ± 0	−40 ± 8	−17	−3	−4 ± 4
	$\bar{q}_L$	0·87	0·95	0·99	1·00	0·86	0·94	0·99	1·00	0·87	0·96	0·99	1·00
$n = 10, a/\sigma = 0{\cdot}35$													
0·1	2	32 ± 23	13	−8	−20 ± 8	6 ± 21	−10	−22	−32 ± 7	19 ± 24	−1	−10	−13 ± 10
	4	−15 ± 18	−9	−4	−5 ± 9	−24 ± 21	−21	−14	−14 ± 7	−12 ± 22	−7	0	2 ± 11
	8	−63 ± 10	−29	−10	4 ± 8	−52 ± 9	−30	−12	−5 ± 6	−60 ± 6	−25	−5	−3 ± 11
	$\bar{q}_L$	0·18	0·28	0·43	0·59	0·18	0·27	0·40	0·55	0·17	0·24	0·28	0·29

Table 2. *Effect of time of crossing in a single cycle structure with an additive model*

($Q = 40$, $M = 8$, $N/M = 2/5$, $n = 5$, $a/\sigma = 0{\cdot}5$, $c = 0{\cdot}5$. Relative response with approx. S.E. for each entry in column.)

		$q = 0{\cdot}1$		$q = 0{\cdot}2$		$q = 0{\cdot}3$		$q = 0{\cdot}4$	
v	T	$t = 10$	∞	$t = 10$	∞	$t = 10$	∞	$t = 10$	∞
2	1	−26	−28	−9	−16	−9	−2	−2	0
	3	−6	−35	−14	−27	−14	−5	5	0
	7	−4	−31	−18	−19	—	—	—	—
4	1	−12	−6	3	−1	1	0	−1	0
	3	−6	−11	−20	−9	−9	0	−6	0
	7	−18	−13	−36	−7	—	—	—	—
8	1	−25	−4	−15	−1	−11	0	−7	0
	3	−22	−7	−22	−9	−18	0	−9	0
	7	−41	−10	−53	−5	—	—	—	—
	Approx. S.E.	15	7	10	2	4	1	4	0

In Table 1 results are given to show the effect of initial gene frequency for a model with other parameters remaining the same as in Fig. 2. Here, and in later tables, the structures are compared in terms of their *relative response*, R_t. Denoting the initial mean by μ_0, the mean of the large population by L_t and that of the other structure by Y_t at generation t, then

$$R_t = 100(Y_t - L_t)/(L_t - \mu_0).$$

Values of R_t are given at intermediate generations and at the limit ($t \to \infty$). With the lowest gene frequency ($q = 0{\cdot}1$) the results in Table 1 are essentially the same as in Fig. 2 ($q = 0{\cdot}2$) in that intense between-line selection has most effect on the limit when there is free recombination. At the higher gene frequencies shown, the chance of fixation of individual genes in the single population approaches 1·0. Then there is little reduction in the limit with between-line selection, and the mean performance with the single cycle structure may be higher for several generations following crossing. Also included in Table 1 is a model with a low initial frequency, a larger number of loci (10) and smaller gene effects than the other models in the Table. The chance of fixation in population L is now only 0·59 for free recombination and 0·29 for complete linkage. However, the results are very similar to those of the model with five loci and $q = 0{\cdot}1$ or 0·2.

(*b*) *Length of the first cycle.* In Table 2 comparisons are made of alternative times (T) of selection between lines (after 1, 3 or 7 generations in sublines) using the same models as in Fig. 2 and Table 1. Only free recombination is included since greater differences are likely to be found than with linkage. In these results, and others not shown, we find that the limit is scarcely and inconsistently affected by the time of crossing, since sampling errors are large relative to the differences we observe. The time of crossing does, of course, affect the mean at intermediate generations (Table 2). When all sublines are chosen the line cross mean is higher at generation 10

Table 3. *Effect of number of sublines in a single cycle structure for an additive model with* $T = 3$, $c = 0{\cdot}5$

(Relative response, or mean gene frequency in $L(\bar{q}_L)$.)

						t			
n	q	a/σ	Q	v/m	N/M	5	10	20	∞
5	0·2	0·5	40	4/8	2/5	−24	−20	−11	−9
				2/4	4/10	3	−7	−8	−9
					$\bar{q}_L$	0·42	0·61	0·85	0·93
10	0·1	0·5	80	4/16	2/5	3	−6	−11	−15
				2/8	4/10	18	−5	−16	−22
				8/16	2/5	−23	−18	−8	−5
				4/8	4/10	−3	−2	−8	−7
				2/4	8/20	−3	−2	−2	−3
					$\bar{q}_L$	0·23	0·40	0·72	0·81
10†	0·5	0·1	80	4/8	2/10	−46	−18	−3	−2
				2/4	4/10	−41	−24	−19	−11
					$\bar{q}_L$	0·58	0·65	0·73	0·88

† Simulation terminated at $t = 80$.

if the crossing is made early since no use is made of the between-line variance. However, with intense between-line selection, temporarily higher means may be obtained with later between-line selection since a larger selection differential can be attained as the variance between lines increases with drift.

(*c*) *Number of sublines.* If the total facilities are kept constant, an increase in the number of sublines must be accompanied by a decrease in the size of each. Thus, at a given time, the variance within lines is reduced and that between lines increased, so the relative efficiencies of within-line and between-line selection may be altered. Results for several models are given in Table 3, each for free recombination. When no selection is practised between lines the limit is independent of the number of sublines (Maruyama, 1970) and no results are included in the table. However, even when selection is practised between the lines, the effect of changing the number of sublines on the limit is small and not significant if the proportion selected within and between lines is not altered. There is one exception in Table 3: $v/m = 2/8$ is poorer than 4/16 for a model with low initial frequency and $a/\sigma = 0{\cdot}5$. However, both schemes are poorer than the single population. At intermediate generations the number of sublines has more effect; higher responses are obtained when the size of the individual sublines is increased.

(*d*) *Total size of the programme.* The relative efficiency of the single cycle and large population structures are compared in Table 4 for different total population sizes (Q). In both schemes the chances of fixation are, of course, increased at larger Q values since the same within-line selection intensities are used. Therefore, although we find smaller differences between the structures at the higher Q values, this is probably solely because the probabilities of fixation approach unity, and we have the same effect as with increase in initial frequency (Table 1). But from the practical

Table 4. *Effect of total number recorded in a single cycle structure with an additive model*

($c = 0{\cdot}5$, $m = 8$, $T = 3$ and $N/M = 0{\cdot}4$. Relative response $\pm$S.E. or mean frequency in $L(\bar{q}_L)$.)

				t				t			
n	q	a/σ	v	5	10	20	∞	5	10	20	∞
				$Q = 40$				$Q = 80$			
5	0·1	0·5	2	6 ± 19	-6	-27	-35 ± 4	0 ± 12	-12	-22	-27 ± 5
			4	-18 ± 14	-6	-6	-11 ± 6	-32 ± 9	-24	-16	-12 ± 5
			8	-50 ± 10	-22	-4	-7 ± 5	-48 ± 7	-23	-8	0 ± 3
			$\bar{q}_L$	0·22	0·36	0·60	0·73	0·23	0·43	0·77	0·92
				$Q = 80$				$Q = 160$			
10	0·1	0·5	2	18 ± 10	-5	-16	-22 ± 3	21 ± 10	6	-3	-7 ± 2
			4	-3 ± 10	-2	-8	-7 ± 3	-11 ± 5	-11	-5	-1 ± 1
			8	-29 ± 11	-14	-2	4 ± 3	-33 ± 4	-19	-6	-1 ± 1
			$\bar{q}_L$	0·22	0·40	0·72	0·81	0·23	0·42	0·78	1·00
				$Q = 80$				$Q = 160$			
10†	0·5	0·1	2	10 ± 19	1	2	-3 ± 4	11 ± 15	3	4	-1 ± 2
			4	7 ± 15	1	7	5 ± 3	-23 ± 19	-17	-1	1 ± 2
			$\bar{q}_L$	0·56	0·61	0·69	0·93	0·57	0·62	0·71	0·96

† Simulation terminated after 80 generations.

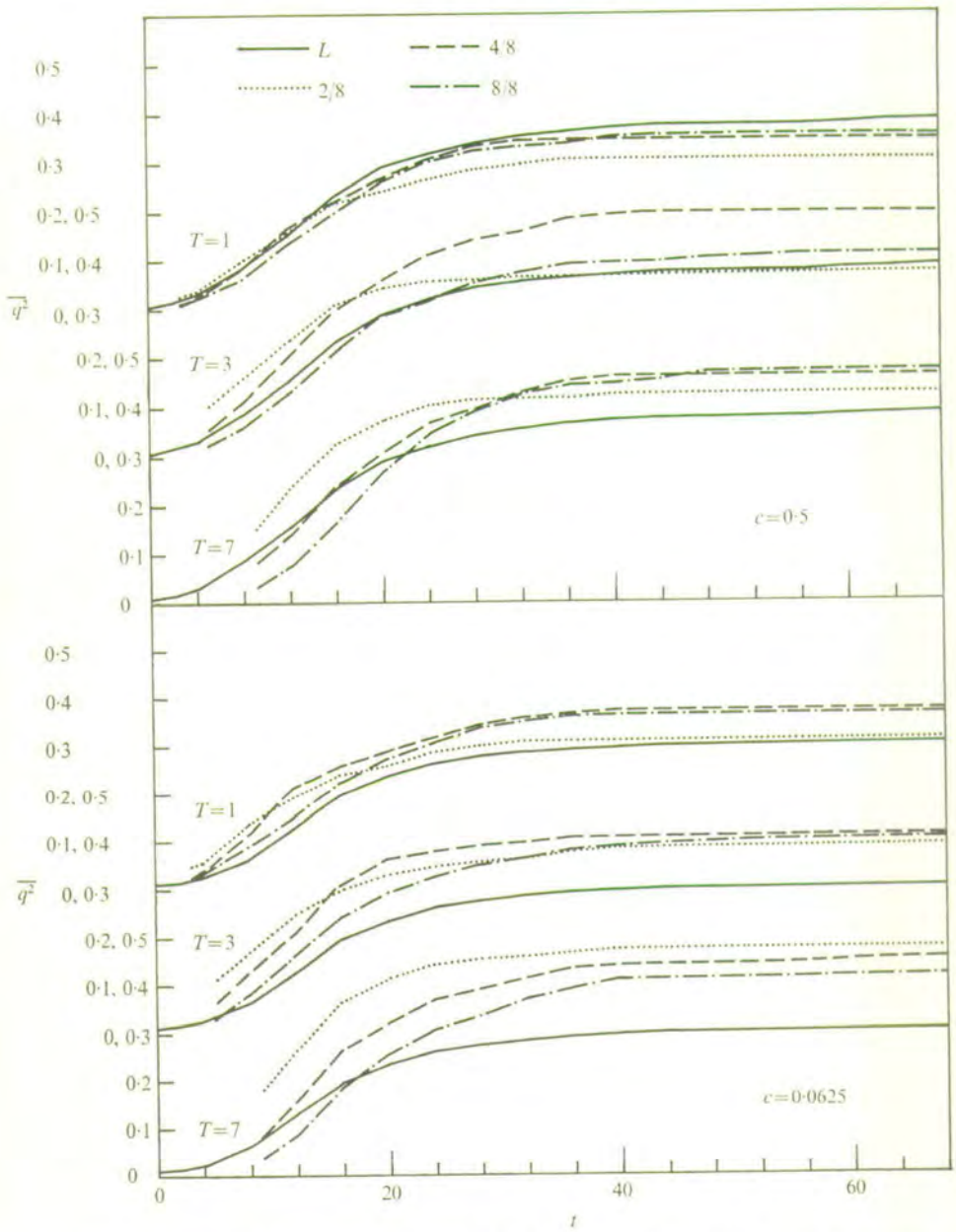

Fig. 3. Comparison between selection in a large population (L) and alternative intensities of between-line selection in the single cycle structure with $Q = 40$, $m = 8$, $M = 5$, $N = 2$ for a recessive model with $n = 5$, $a/\sigma = 0{\cdot}5$, $q = 0{\cdot}1$ and recombination fraction, c, for three times (T) of crossing after subdivision. The population mean is a function of $\overline{q^2}$.

viewpoint this is important, since we have schemes where the mean of the single cycle structure exceeds that of the single populations for a long period with little sacrifice at the limit – for example, when $q = 0{\cdot}1$, $a/\sigma = 0{\cdot}5$, $v = 2$ and $Q = 160$ (Table 4).

Table 5. *Effects of selection intensity between lines, initial gene frequency and linkage in a single-cycle structure with a recessive model*

($Q = 40$, $m = 8$, $N/M = 2/5$, $T = 3$, $n = 5$, $a/\sigma = 0{\cdot}5$. Relative response ± S.E., or mean performance of $L(\bar{q}_L^2)$.)

		$c = 0{\cdot}5$				$c = 0{\cdot}0625$				$c = 0{\cdot}0$			
q	v	$t = 5$	10	20	∞	$t = 5$	10	20	∞	$t = 5$	10	20	∞
0·1	2	201 ± 66	62	18	−4 ± 11	361 ± 89	154	44	29 ± 12	170 ± 96	70	6	4 ± 11
	4	46 ± 29	36	26	30 ± 12	141 ± 45	78	58	37 ± 14	42 ± 65	11	1	7 ± 11
	8	−41 ± 19	−33	−1	6 ± 8	−13 ± 20	25	26	34 ± 12	−34 ± 23	−11	−5	−4 ± 13
	$\bar{q}_L^2$	0·04	0·12	0·29	0·38	0·03	0·10	0·23	0·30	0·05	0·13	0·28	0·30
0·4	2	0 ± 7	−20	−11	−9 ± 3	28 ± 12	−6	−12	−11 ± 4	45 ± 11	−16	5	1 ± 5
	4	−38 ± 7	−15	−6	−6 ± 3	−21 ± 7	−19	−8	−9 ± 3	−5 ± 8	14	13	8 ± 4
	8	−69 ± 6	−32	−11	−1 ± 2	−72 ± 6	−40	−20	−5 ± 4	−51 ± 8	−9	7	5 ± 5
	$\bar{q}_L^2$	0·48	0·74	0·91	0·94	0·44	0·68	0·87	0·90	0·44	0·65	0·73	0·76

Table 6. *Effect of intensity and time of between line selection and initial gene frequency in a single-cycle structure with a recessive model*

($Q = 40$, $m = 8$, $N/M = 2/5$, $n = 5$, $a/\sigma = 0{\cdot}5$, $c = 0{\cdot}5$. Relative response, or mean performance of $L(\bar{q}_L^2)$.)

		$q = 0{\cdot}1$				$q = 0{\cdot}4$				$q = 0{\cdot}7$			
v	T	$t = 5$	10	20	∞	$t = 5$	10	20	∞	$t = 5$	10	20	∞
2	1	49	0	−18	−21	−12	−15**	−9**	−5	2	−3	1	0
	3	201**	62**	18	−4	0	−20**	−11**	−9*	−3	−4	−1	−2
	7	—	54*	30	11	—	−22**	−8	−4	—	−10**	−4	−3
4	1	−8	2	−9	−11	−26**	−6	0	1	−4	−2	0	0
	3	46	26	26	30*	−38**	−15*	−6	−6	−24**	−3	0	0
	7	—	−25	5	19	—	−4	1	3	—	−23**	0	0
8	1	−33	−21	−9	−9	−43**	−20**	−2	2	−24**	−8**	−2	−2
	3	−41	−33	−1	6	−69**	−32**	−11**	1	−48**	−16**	−3	−2
	7	—	−72**	−7	21	—	−38**	−4	5*	—	−40**	−1	0
	$\bar{q}_L^2$	0·04	0·12	0·29	0·38	0·48	0·74	0·91	0·94	0·85	0·97	1·00	1·00

* $0{\cdot}01 < P < 0{\cdot}05$; ** $P < 0{\cdot}01$ relative to L.

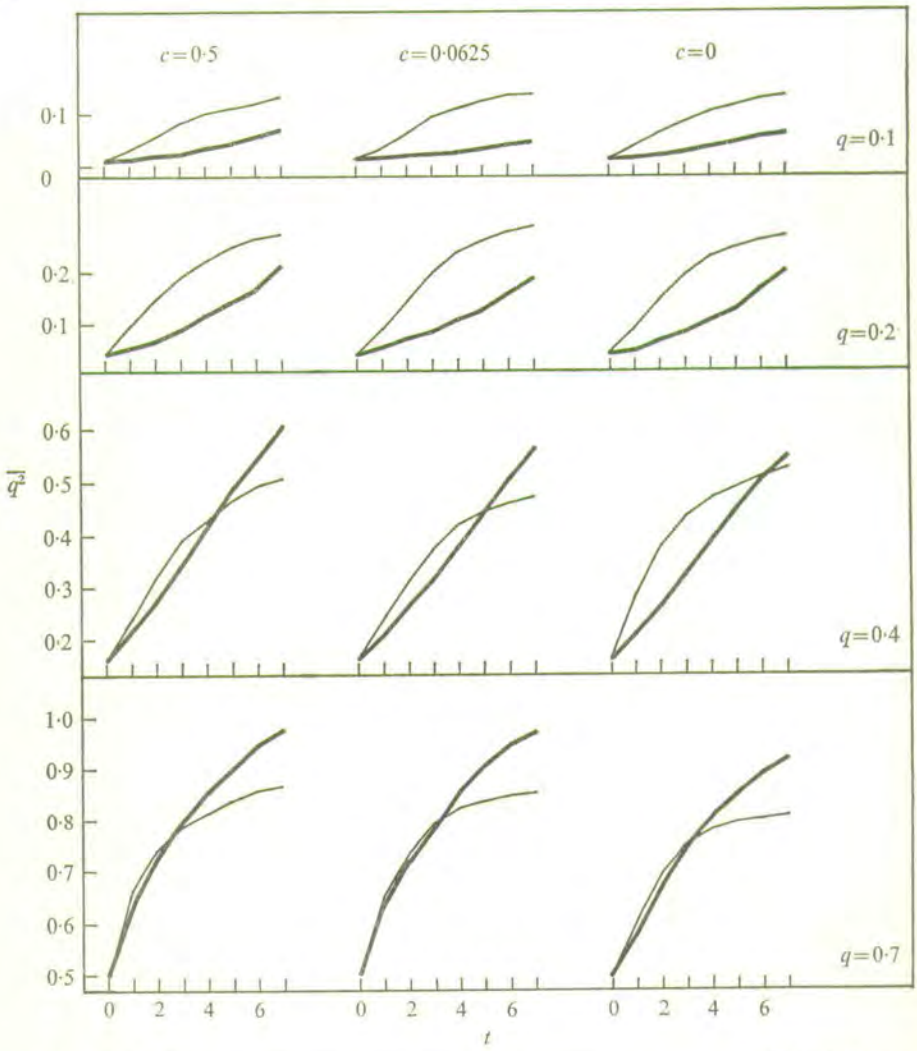

Fig. 4. Comparison between the mean performance (expressed as $\overline{q^2}$) of the large population (thick lines) and the mean of the sublines (thin lines) prior to crossing, for a recessive model with $n = 5$, $a/\sigma = 0{\cdot}5$ and specified initial frequencies (q) and recombination fraction (c). The numbers selected/recorded are 16/40 in the large population and 2/5 in the sublines.

(ii) *Single-cycle structure: recessive model*

We shall use the term 'recessive model' when, at each locus, there is complete dominance and the recessive allele is favoured by selection. If all loci have the same effect on the quantitative trait, the mean performance is a linear function of $\overline{q^2}$, where q is the gene frequency at a single locus in a single replicate. This statistic is used in the figures and tables.

Results for recessive models are given in Tables 5 and 6 and Fig. 3. The structures used are similar to those investigated earlier for the additive model, but the results differ considerably. We find that immediately following between-line selection the mean may be higher than in the single population and can remain ahead at the

limit. These effects are seen most markedly with low initial gene frequencies such as 0·1 (Fig. 3). At higher initial frequencies, such as 0·7 (Table 6) when all favourable alleles are fixed, the mean of the single cycle structure does not exceed that of the single population and the same limit is reached. In general we see that the different intensities of between-line selection have rather small effect on the limit, but, of course, large effects at intermediate generations.

In the recessive model the length of the cycle of sublining has an important influence on the limit. We see in Table 6 that where the schemes differ appreciably in efficiency at low initial frequencies, the highest limits are attained when the between-line selection is delayed. But if no between-line selection is practised the intermediate generations are poorer when crossing is delayed for the lines have ceased to respond to within-line selection. With very tight linkage we find, as in the additive model, that the different intensities of between-line selection do not influence the limit markedly (Table 5).

In Fig. 4 the responses in the initial generations of sublines are compared with those of the single population. In contrast with the additive model, higher rates of gain may be made in the very small lines if the initial frequency is low. In these situations the additive variance actually increases up to intermediate levels of inbreeding (Robertson, 1952). In addition, when the recessive alleles are favoured, there is an inbreeding, enhancement' as homozygotic frequency increases. This is lost in crossing and we see (Fig. 3) that with no between-line selection the line cross is at first poorer than the single population.

(iii) *Single-cycle structure: dominant models*

Some results are given in Table 7 for a model of equal effects and initial gene frequencies with free recombination, in which there is complete dominance with the dominant allele favoured by selection. If it has a low initial frequency the response is less in the single-cycle structure than in the large population throughout the selection period. However, at the limit the difference is small if no between-line selection is practised. In addition, prior to crossing, the sublines perform much more poorly than the single population since the lines exhibit inbreeding depression. At higher initial frequencies of the dominant allele the pattern alters, for as we have seen in the previous section the efficiency of within and subsequently between-line selection is enhanced if the lines are small. However, in our example the chance of fixation is very high and only small differences are observed at the limit. We consider these models further in the repeated cycle scheme.

(iv) *Repeated-cycle structure*

All repeated cycle studies were undertaken with the intermediate cycle length $T = 3$. A typical run with an additive model is shown in Fig. 5, in which the parameters are the same as those used in Fig. 2, and further results are given in Table 8. In each case comparison is made with the large population system.

The repeated subdivision with no between-line selection gives essentially the same limit as the single population (or single cycle) structure, but the limit is reached

Table 7. *Effect of intensity and time of selection between lines in a single cycle structure with a dominant model*

($Q = 40$, $m = 8$, $N/M = 2/5$, $n = 5$, $a/\sigma = 0{\cdot}5$, $c = 0{\cdot}5$. Relative response, or mean performance of L expressed as $1-\overline{(1-q_L)^2}$.)

		$q = 0{\cdot}1$				$q = 0{\cdot}4$				$q = 0{\cdot}7$			
v	T	$t = 5$	10	20	∞	$t = 5$	10	20	∞	$t = 5$	10	20	∞
2	1	−12	−25**	−19**	−30**	−2	−3	1	0	−2	−7	6	2
	3	−21*	−35**	−38**	−39**	−22*	−5	−3	−5	−27	9	11	3
	7	—	−39**	−39**	−38**	—	−18**	−4	−5	—	16	8	3
4	1	−5	−4	−5	−7	1	−1	2	−1	21	4	2	3
	3	−29**	−20**	−14**	−14**	−26**	−6	−1	0	−4	−7	0	3
	7	—	−42**	−22**	−18**	—	−14*	1	0	—	12	6	3
8	1	−27**	−11	0	0	−10	−2	1	0	8	4	4	3
	3	−59**	−24*	−12	−10	−36**	−10**	−1	0	−39	−10	−13	2
	7	—	−55**	−16**	−7	—	−29**	−4	0	—	−34	−11	3
$1-\overline{(1-q_L)^2}$		0·49	0·70	0·84	0·90	0·84	0·92	0·97	1·00	0·94	0·96	0·99	1·00

* $P < 0{\cdot}05$, **$P < 0{\cdot}01$ of zero relative response.

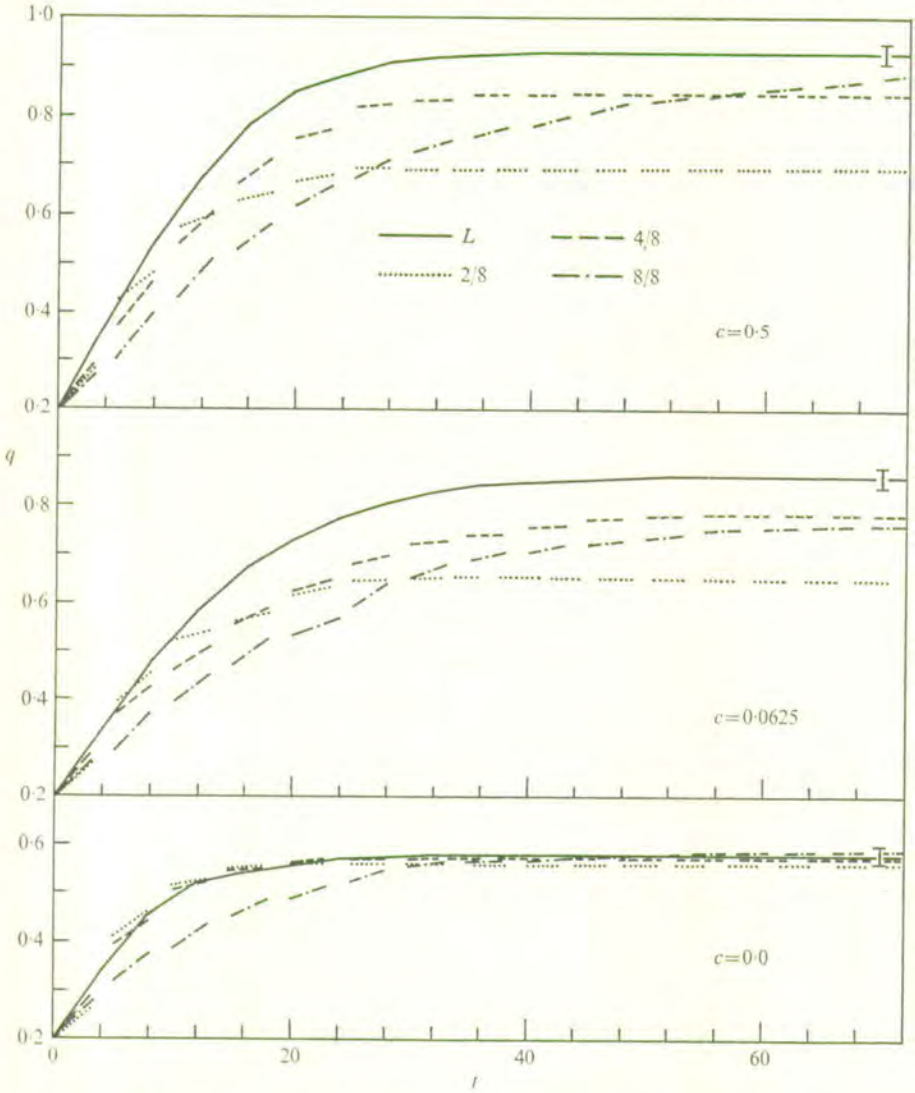

Fig. 5. Comparison between selection in a large population (L) and alternative intensities of between line selection in a repeated cycle structure with an additive model as for Fig. 2: $Q = 40$, $m = 8$, $M = 5$, $N = 2$, $n = 5$, $q = 0{\cdot}2$, $a/\sigma = 0{\cdot}5$.

at a much slower rate. There are, of course, a large number of generations in which no within-line selection is practised following each cross and these both reduce the rate of advance and also the limit to a small extent. With intense selection between the lines the rate of advance is increased, such that in the example shown in Fig. 5 when $v/m = 2/8$ and linkage is complete, the repeated cycle is superior to the large population for the greater part of the two cycles after first crossing, and finally a similar limit is reached. However, with free recombination or partial linkage, the response soon drops below that of the large population, and a lower limit is attained. It is clear that the single cycle and multiple cycle schemes give essentially the same results.

Table 8. *Repeated cycle structure with an additive model*

($T = 3$, $c = 0{\cdot}5$. Response relative to L.)

							t			
Q	m	N/M	n	a/σ	q	v	5	10	20	∞
40	8	2/5	5	0·5	0·1	2	18	3	−25	−36
						4	4	6	−5	−11
						8	−33	−39	−25	−2
					0·5	4	−24	−16	−4	0
					0·7	4	2	−1	−2	0
80	16	2/5	5	0·5	0·1	2	52	13	−28	−38
						4	26	13	−8	−15
						8	−8	−15	−11	1
						16	−61	−64	−57	−9†
80	8	2/10	10	0·1	0·5	4	−13	−18	−11	−9†
						8	−31	−37	−33	−18†

† Simulation terminated prior to fixation (after 80 generations).

Table 9. *Repeated cycle structure with a recessive and dominant model*

($Q = 40$, $T = 3$, $c = 0{\cdot}5$. Response relative to L.)

				Recessive				Dominant			
a/σ	n	v	q	$t = 5$	10	20	∞†	$t = 5$	10	20	∞†
				$m = 8$, $N/M = 2/5$							
0·5	5	4	0·1	34	42	23	6	−27	−24	−15	−17
			0·7	−11	−4	−2	0	−10	19	9	3
0·35	10	4	0·1	63	48	24	24	−41	−38	−36	−29
			0·4	−4	−8	−6	−1	−24	−20	−12	−5
				$m = 20$, $N/M = 2/2$							
0·5	5	5	0·1	−25	−35	−42§	4	—	—	—	—
		10	0·1	−41	−58	−60§	53‡	—	—	—	—

† $t = 60$ for $a/\sigma = 0{\cdot}35$.
‡ $t = 100$.
§ At $t = 60$ relative response is $+2$ for $v = 5$, $+10$ for $v = 10$.

A few results for non-additive models with repeated cycles are given in Table 9. With the recessive allele initially at low frequency, whether at a selective advantage of disadvantage, greater advances may be made both in the early generations and at the limit. Table 9 also includes a model with a low-frequency-favoured recessive in which no selection is practised within sublines, but with 5/20 or 10/20 sublines selected after $T = 3$ generations each cycle. The rate of advance is very slow, but a much higher limit is reached with the less intense between-line selection scheme than with the large population control system.

When the dominant allele is favoured the pattern of response is very irregular with the repeated cycle scheme, since there are intermittent periods of inbreeding

followed by line crossing to restore heterozygosis. Only after several generations do the sublines become fixed sufficiently for their performance not to fall below that of the single population before they are crossed.

4. DISCUSSION

We have studied a restricted range of genetic models with a rather small number of genes of equal effect and initial frequency, and we must ask whether we are entitled to generalize beyond them. We may be justified in doing so if it can be explained why the alternative schemes performed in the way they did. Most of the discussion will be restricted to additive models, for which the theory has been developed furthest.

(i) *Additive genes*

The important item of existing theory is that any subdivision structure, including one of no subdivision, in a single locus additive model gives the same limit so long as there is no between-line selection and the selection intensity is the same in each population (Maruyama, 1970). Our results show that this generalization holds for multiple loci which recombine freely. Now when selection is practised between lines the mean level of inbreeding in the subsequent single population or second cycle sublines is increased and, at least for an additive model, the genetic variance correspondingly reduced. If the inbreeding level in each subline is F_{T+1} at the generation the crosses are made, then the cross of v lines has inbreeding coefficient F_{t+1}/v. For example, with $N = 2$, $T = 3$ and random mating, $F_4/v = 34{\cdot}2\,\%$, $17{\cdot}1\,\%$ and $8{\cdot}5\,\%$ for $v = 2$, 4 or 8. It is clear from our results that the gain from between-line selection is more than compensated by a reduction in subsequent response. This simple argument can be quantified for an additive model with a large number of independent loci each with genes of small effect, as we now show.

Let us assume that the variances change in proportion to the level of inbreeding, since the populations are mated at random and the mean changes in gene frequency are small (Robertson, 1960). Let the heritability of the trait be h^2 and the phenotypic variance σ_p^2. The response to selection with lines in the first cycle, including the selection within each line for crossing, is

$$\begin{aligned}\mu_{T+1} - \mu_0 &= \sum_{t=0}^{T} (1 - 1/2N)^t i h^2 \sigma_p \\ &= 2NiF_{T+1} h^2 \sigma_p, \end{aligned} \tag{1}$$

where i is the standardized within-line selection differential (which we shall assume depends only on the proportion selected, although it is also marginally affected by the total number scored). The genetic variance between lines at generation T when selection is practised between lines is $2F_T h^2 \sigma_p^2$. The within-line phenotypic variance is then

$$[(1 - F_T) h^2 + 1 - h^2] \sigma_p^2,$$

so if M individuals are recorded, the variance of an observed line mean is

$$2F_T h^2 \sigma_p^2 + [(1 - F_T) h^2 + 1 - h^2] \sigma_p^2 / M,$$

where, as in our simulation model, we assume there is no environmental variance common to all members of a line. Thus with a standardized selection differential of i_B, the response, B, to between-line selection is expected to be

$$B = 2i_B F_T h^2 \sigma_p^2 \{2F_T h^2 \sigma_p^2 + [(1-F_T)h^2 + 1 - h^2]\sigma_p^2/M\}^{-\frac{1}{2}}$$
$$= 2i_B h^2 \sigma_p F_T \{M/[(2M-1)F_T h^2 + 1]\}^{\frac{1}{2}}.$$

In the first t^* generations of within-line selection in a population of size Nm in a single cycle structure, subsequent to crossing and random mating (in a sufficiently large population that drift can be ignored at generation $T+1$), the response is

$$\mu_{T+2+t^*} - \mu_{T+1} = 2NmiF_{t^*} h^2 \sigma_p (1 - F_{T+1}/v),$$

where $F_{t^*} = 1-(1-1/2mN)^{t^*}$ is the inbreeding level relative to that after crossing. Thus the total advance from t^* generations of selection after crossing is

$$\mu_t - \mu_0 = 2h^2\sigma_p[iNF_{T+1} + iNmF_{t^*}(1-F_{T+1}/v) + i_B F_T\{M/[(2M-1)F_T h^2 + 1]\}^{\frac{1}{2}}]$$

and as $t \to \infty$, the limit is

$$\mu_\infty - \mu_0 = 2h^2\sigma_p(iNm - iNF_{T+1}[(m/v)-1] + i_B F_T\{M/[(2M-1)F_T h^2 + 1]\}^{\frac{1}{2}}).$$

If there is no between-line selection, i.e. $v = m$ and $i_B = 0$, then $\mu_\infty - \mu_0 = 2Nmih^2\sigma_p$, which is the total advance expected in the large population (L) without any subdivision, with this simple model in which the genetic variance is directly proportional to the level of inbreeding.

Using the above formulae we have calculated the advance for the structures used in our simulation studies, and have assumed that $h^2 = 0{\cdot}2$ and line means are normally distributed. This heritability is slightly larger than those used in the simulation (e.g. Table 1, Fig. 2). The results are shown in Fig. 6, using two different scales for time: either generations (t) or $F = 1-(1-1/32)^t$, which is the inbreeding coefficient in L at generation t. On the latter scale the responses in both L and the other large populations after crossing of sublines are linear. Since these results strongly resemble those obtained earlier for additive models with free recombination (Fig. 2), they illustrate the utility of the simple model. Only when between-line selection is practised early and is intense does the response in the single-cycle structure exceed that in the large single population, but then the limit is reduced. The limit is least affected when between-line selection and crossing is done as early as possible, thereby minimising inbreeding in the subsequent population. However, with early crossing less response is made directly from the between-line selection. In our simulation studies we were unable to detect which effect was larger, but presumably would have shown that short cycles of inbreeding gave the highest limits if sufficient replicate computer runs had been made. With cycles of length of only one generation the repeated cycle structure degenerates into a family selection scheme, and since with comparable selection intensities family selection gives a lower limit than mass selection (Robertson, 1960) our results could be anticipated.

A less precise argument on the effects of between-line selection can be used and then extended to include linkage. Imagine the trait under selection is controlled by a few, say 8, independent genes of low initial frequency and that selection is continued in sublines until all the loci are fixed, with the probability of fixation

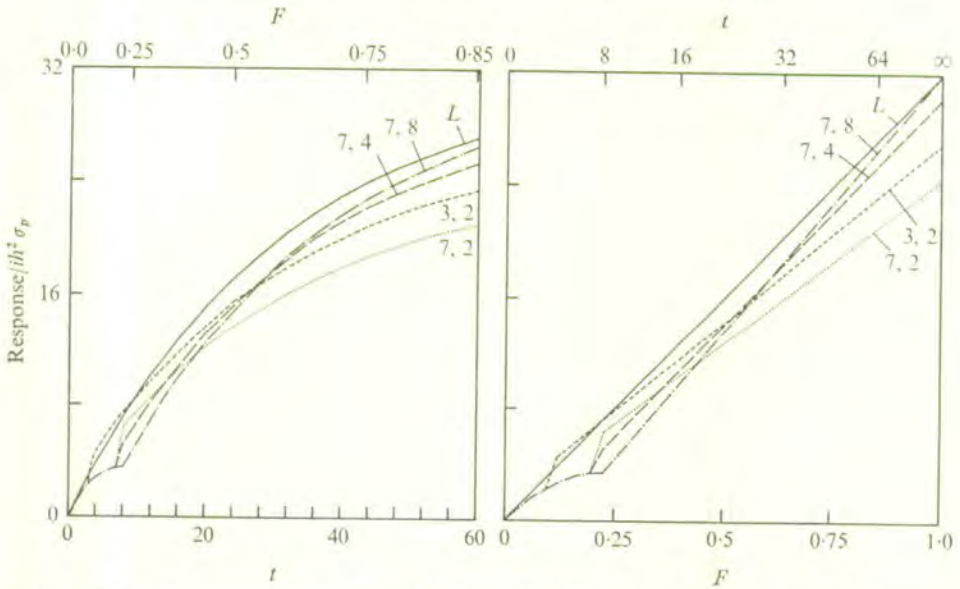

Fig. 6. Responses predicted for an additive model of small gene effects in a single large population (L) or in a single cycle structure with (T, v) generations of inbreeding and lines selected, and $Q = 40$, $m = 8$, $M = 5$, $N = 2$, $h^2 = 0{\cdot}2$. The response is shown as the coefficient of $ih^2\sigma_p$, where i is the within-line selection intensity and σ_p the phenotypic standard deviation for two time-scales: generations, t, and inbreeding coefficient in L, $F = 1-(1-1/32)^t$.

of the favourable allele being 0·25 at each locus. Thus the probability that any line contains 0, 1, 2, 3, 4, $\geqslant 5$ favourable alleles at fixation is 0·10, 0·27, 0·31, 0·21, 0·09, 0·03 respectively, from the binomial distribution. Imagine also that there are eight sublines, so this is also the frequency distribution of the number of sublines which contain the favourable allele at a specified locus. When no selection between lines is practised, there is thus a 90 % chance of having at least one favourable allele at this locus, which, with an initial frequency of at least 1/8, would have a fairly high chance of fixation in the new, larger, population. By contrast, imagine only the best two sublines are chosen. The probability that a line contains at least 4 favourable alleles is 0·12 (or, more precisely, 0·1138), so the probability that at least 2 of 8 lines have 4 or more favourable alleles is $1-(0{\cdot}88)^8-8\times 0{\cdot}12\times(0{\cdot}88)^7 = 0{\cdot}23$. Thus, even if the two sublines were chosen without error, in only 23 % of samples would these both contain 4 or more favourable alleles, and even if both contain 4 favourable alleles, the probability that the allele at a specific locus is present is only 75 %.

With free recombination the crucial requirement is that at least one representative of the favourable allele at each locus should occur in the cross of selected lines, for subsequent recombination will permit formation of the best possible chromosomes. At the other extreme, if all genes affecting the trait under selection are completely linked on a single chromosome, the most desirable outcome is to retain the best chromosome, initially sampled at the start of the experiment in one subline, during selection between lines and subsequent selection. Now since the between-

line selection has a high accuracy, the line containing the best chromosome has a high chance of being selected, even if only two or so lines are chosen. (Even if it is missed, the next best chromosome will probably be chosen.) Thus the probability of fixing the best chromosome should be little affected by the intensity of between-line selection, and this is the result we obtain. Further, we do not expect to find large differences between sublined structures and a single large population when linkage is complete since the best, or nearly the best, initial chromosome is fixed in either case. The relevance of this kind of genetic model in selection limits in single populations is discussed further by Robertson (1970*a*).

Of course, in nature we have neither independent loci nor complete linkage on single chromosomes, but a mixture of linkage relationships on individual chromosomes together with independence of genes from different chromosomes. Our results show that, with some recombination, the selection between lines has an effect intermediate between that of independence and complete linkage. Thus even for species with few chromosomes we must expect that selection between sublines in the structures we have considered could markedly reduce the limit if the trait is affected mostly by additive genes.

We have undertaken a small number of computer runs with the restriction of equal gene effects or frequencies removed. Using the same structural parameters as in Figure 2, a model was simulated of five additive loci with equal initial frequency and effects $a/\sigma = 0{\cdot}875, 0{\cdot}5, 0{\cdot}375, 0{\cdot}25$ and $0{\cdot}177$, such that the genetic variance is the same as in a model of five loci of effect $a/\sigma = 0{\cdot}5$. The general pattern was found to be similar to that of equal effects, but between-line selection had rather less effect at the limit, presumably because those genes with the largest effect have a high chance of being selected and these contribute most to the total advance. With a more extreme additive model of one locus with $a/\sigma = 1$ and $q = 0{\cdot}025$ and nine loci of $a/\sigma = 0{\cdot}25$ and $q = 0{\cdot}4$, the probability of fixing the gene of large effect was little influenced by the structure, whereas between-line selection reduced the probability of fixing those genes of smaller effect.

In an attempt to utilize the immediate response from selection between lines but to minimize the somewhat drastic effects of truncation selection between lines on the limit we tested a scheme whereby a high proportion of chromosomes to form the line cross pool were taken from the best lines, but some were allowed to enter from the poorer ones. However, we were not successful in this attempt: in order to attain large gains from between-line selection the limit had to be sacrificed.

(ii) *Intermediate generations*

We have concentrated our attention on the mean performance and selection limits after crossing the replicate sublines. However in practice it might be possible to utilize the variation between the sublines by choosing one for multiplication and commercial use, if only on a temporary basis. Baker & Curnow (1969) have estimated this variance between lines for a range of genetic models and shown that the best sublines are likely to be very superior to a large contemporaneous population. Using the model of small gene effects described above, A. Robertson (personal com-

munication) has derived formulae for the relative merits of the best subline and a single large population. He has kindly let us present his analysis, which is based on some approximations appropriate for sizes of sublines rather larger than those used in this study (say $N \geqslant 8$). From equation (1), the expected gain in the sublines of size N after t generations is

$$\begin{aligned} \mu_t - \mu_0 &= 2Nih^2\sigma_p[1-(1-1/2N)^t] \\ &= ih\sigma_A(t-t^2/4N) \quad \text{approximately,} \end{aligned} \tag{2}$$

where σ_A^2 is the additive variance, and provided N is not too small. Thus the reduction in response due to inbreeding up to the tth generation, relative to using a very large population in which inbreeding effects are negligible in this period, is $ih\sigma_A t^2/4N$. At the same time, the genetic variance between lines will be $2F_t\sigma_A^2 = t\sigma_A^2/N$ approximately. If the expected superiority of the best line of the set is k times the standard deviation between them (i.e. $k = i_B$ when one line is chosen), the expected superiority of the best line over the large population may be written as

$$D = \sigma_A[-iht^2/4N + k\sqrt{(t/N)}], \tag{3}$$

which passes through a maximum when $t^3 = Nk^2/i^2h^2$, giving $D = (k^4/Nih)^{1/3}3\sigma_A/4$.

In the selection experiment with *Drosophila melanogaster* of Madalena (1970) there were eight sublines of $N = 10$, with $ih = 0{\cdot}8$. Then the greatest difference between the best subline and the large population is expected at generation 3 when $D = 0{\cdot}58\sigma_A$, about 25 % of the response in the large population at that time. The actual difference was smaller, but could be explained by sampling. In our example of Fig. 2 we have $N = 2$, $i = 1$, $k = 1$, $h^2 = 0{\cdot}1$, approximately, and the above formulae predict that $t_{\max}$ lies between 3 and 4, and that at generation 3, $D = 1{\cdot}3\sigma_A$, whereas in the large population at this time the response would be $0{\cdot}9\sigma_A$. Although formulae such as (2) and (3) do not hold exactly in our example, since N is so small, the prediction is essentially correct for direct calculation gives a maximum D of $1{\cdot}35\sigma_A$ at generation 4. As Baker & Curnow (1969) have shown numerically, the best subline is likely to be much superior to a large population for only a short time. Eventually the large population is likely to be best.

However, the above analysis requires that the line of best genotype be identified. In our simulation experiments small samples were measured each generation, but accuracy of choosing lines could have been improved by recording line means for several generations. For example, the correlation of line means for the model of Fig. 2 was only 0·36 between generations 3 and 7. In addition, these gains from selecting the best line last only a few generations, and although a consequent loss at the limit need not be incurred since all lines can be crossed, the mean of crosses is then poorer than that of a single population selected throughout. There is probably need for further study of methods of structuring populations to make the best use of short-term benefits.

All our comparisons have been made at the same selection intensity within sublines and the large population. Higher response in the initial generation, at the expense of the limit, can be obtained by selecting more intensely within the large

population. This is a much simpler scheme, and can give essentially the same results as a period of subdivision followed by between-line selection and crossing.

(iii) *Dominant genes*

When there is dominance we have seen that the effects of subdivision and between-line selection may differ markedly from those with additive genes. Firstly there is an *increase* in the additive variance in small random mating populations if the recessive alleles are at low frequency (Robertson, 1952). Therefore, as we have seen in Fig. 4, the response in the cycle of subdivision may be higher in the sublines than in the large population. More important, perhaps, the variance between lines at fixation is a function of $q(1-q)$, whereas the initial additive variance is proportional to $q^3(1-q)$ (where q is the frequency of the recessive allele, which is not assumed to change much during selection, i.e. we adopt a small effects model for illustration). Thus the between-line variance and response can be of a different order of magnitude to that within a single large population if the recessive allele is at low frequency. The between-line variance increases in proportion to F^3, where F is the inbreeding coefficient, so it becomes much more efficient if between-line selection is delayed, as our simulations results show. At these later times both the single and repeated cycle schemes give higher responses both in intermediate generations and at the limit than does the single large population. However, we see from our results that intense between-line selection depresses the limit (at least below that for no selection between lines in the same structure) for the same reasons as given in the additive model, and that favourable alleles at some loci are lost during this restriction of population size.

When the recessive alleles are at intermediate or high frequency we have found that a structured scheme is not of benefit, and, as predicted in the additive case, delaying between-line selection gives lower responses in intermediate generations, as well as lower limits. The arguments of the previous section on low-frequency recessives now act in reverse. We have simulated some models with both additive and completely dominant genes (with the recessive favoured) and found that between-line selection influences response in a manner roughly intermediate between that for additive and recessive models taken separately.

Few, if any, quantitative traits of economic importance show negative heterosis. Therefore it is unlikely that much useful variation is expressed at loci in which the recessive alleles are favoured, so we can suggest that the kind of structured systems discussed here are only likely to be useful for removing deleterious recessive genes initially at low frequency. In other genetic situations it seems unlikely that there will be sufficient extra gain in initial generations from between-line selection to compensate for the potential loss at the limit when between-line selection is practised and similar gains can be obtained simply by using more intense selection within single populations. Small and temporary benefits can be obtained, however, from using the best sublines prior to crossing. We have not investigated epistatic models, for which these line crossing systems were originally proposed by Wright, and some studies with these models could be rewarding.

REFERENCES

BAKER, L. H. & CURNOW, R. N. (1968). Choice of population size and use of variation between replicate populations in plant breeding selection programs. *Crop Science* **9**, 555–560.

BOWMAN, J. C. & FALCONER, D. S. (1960). Inbreeding depression and heterosis of litter size in mice. *Genetical Research* **1**, 262–274.

DEMPSTER, E. R. (1955). Genetic models in relation to animal breeding problems. *Biometrics* **11**, 535–536.

GOODWILL, R. (1971). Effect of population structure in selection experiments with *Tribolium castaneum*. *Genetics* **68**, s 24.

HILL, W. G. (1963). Cyclical inbreeding with selection in *Drosophila melanogaster*. Unpublished M.S. thesis. University of California, Davis.

HILL, W. G. & MADALENA, F. E. (1969). Optimum population structure in selection programs. *Genetics* **61**, s 26–27.

HILL, W. G. & ROBERTSON, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

JAMES, J. W. (1972). Optimum selection intensity in breeding programmes. *Animal Production* **14**, 1–9.

MADALENA, F. E. (1970). Studies on the limits to artificial selection. Unpublished Ph.D. thesis, University of Edinburgh.

MARUYAMA, T. (1970). On the fixation probability of mutant genes in a subdivided population. *Genetical Research* **15**, 221–227.

POLLAK, E. (1966). On the survival of a gene in a subdivided population. *Journal of Applied Probability* **3**, 142–155.

ROBERTSON, A. (1952). The effects of inbreeding on the variation due to recessive genes. *Genetics* **37**, 189–207.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London* B **153**, 234–249.

ROBERTSON, A. (1970*a*). A theory of limits in artificial selection with many linked loci. In *Mathematical Topics in Population Genetics* (ed. K. Kojima), pp. 246–288. Berlin: Springer.

ROBERTSON, A. (1970*b*). Some optimum problems in individual selection. *Theoretical Population Biology* **1**, 120–127.

WRIGHT, S. (1939). Genetic principles governing the rate of progress of livestock breeding. *Proceedings of the American Society of Animal Production*, **32**, 18–26.

WRIGHT, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

7

Probability of fixation of genes in populations of variable size

by

William G. Hill

Reprinted from THEORETICAL POPULATION BIOLOGY
All Rights Reserved by Academic Press, New York and London

Vol. 3, No. 1, March 1972
Printed in Belgium

Probability of Fixation of Genes in Populations of Variable Size*

WILLIAM G. HILL†

Statistical Laboratory, Iowa State University, Ames, Iowa 50010

Received September 4, 1970

Natural populations undergo wide fluctuations in size from season to season and from year to year in response to changes in the environment. In many species, therefore, the amount of gene-frequency drift differs from generation to generation, and there must also be differing selective forces according to whether the population is expanding or contracting in size. It seems reasonable to assume that when there is a drastic reduction in population size, say in a severe winter, there are greater relative differences in survival probabilities of different genotypes than under conditions of expanding population in a less hostile environment.

In most studies of the effects of finite population size an assumption of constant size and structure has been made. Wright (1939) and Crow (1954), however, indicated that over a period of several generations the average effective population size was given by the harmonic mean of the population sizes found in this period. Recently an exact treatment for drift with no selection has been given by Karlin (1968). He provides a method for finding the asymptotic approach to homozygosity in general models of distribution of progeny number and changes in population size. Karlin specifies a matrix of transition probabilities of population sizes in successive generations, and we shall use this matrix for a model in which selection is included.

The probability of gene fixation (absorption probability) and rate of approach to homozygosity with selection in populations which change cyclically in size has been discussed by Chia (1968). In particular he shows that if the cycle length is of a smaller order of magnitude than the population sizes in the cycle, each assumed to be of similar order, then a diffusion equation can be used to compute the fixation probabilities. The same result for the case of a single new mutant

* Journal Paper No. J-6601, Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 1669. Supported by National Institutes of Health, Grant No. GM 13827.

† Present address: Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN.

with this cyclical model is given by Ewens (1967) using branching processes and Kimura (1960) using the diffusion equation.

In this paper we shall assume a very general form for changes in population size, and develop a matrix series expansion to obtain absorption probabilities. The method is essentially an extension of one used by Hill (1970) and Narain and Robertson (1969) for populations of constant size. The approximations obtained do not hold when both selection forces are strong and population sizes are large. Under rather different assumptions a diffusion approximation is also developed. A few worked examples are given

Matrix Series Approximation

We assume a haploid model with a single locus having two alleles, A, a with no mutation. There are r possible values for the population size, N_i, all of which are finite, such that the population can neither become extinct, nor increase indefinitely in size. Following Karlin (1968) we can specify a transition probability matrix $\mathbf{C}$ with elements c_{ij}, where

$$c_{ij} = P \text{ (population size is } N_j \text{ at generation } t+1 \mid \text{population size is } N_i \text{ at generation } t)$$

independent of t, and N_i, $N_j = N_1, \ldots, N_r$. In general there is no particular ordering in values of the N_i; it is not necessary that $N_i > N_{i-1}$, for example. In order to illustrate the use of $\mathbf{C}$ Karlin gives the example of cyclical variation in population size. Then, for the appropriate arrangement of the N_i, we have

$$\begin{aligned} c_{i,i+1} &= 1, \ 1 \leqslant i \leqslant r-1, \\ c_{r1} &= 1, \\ c_{ij} &= 0, \text{ otherwise.} \end{aligned}$$

We also assume that for each pair of population sizes N_i, N_j there is an associated selective value s_{ij}, which specifies the selective advantage of the A allele over the a allele during a generation in which the population size changes from N_i to N_j. Thus we define a matrix $\mathbf{S}$ with elements s_{ij} for $1 \leqslant i, j \leqslant r$.

We define

$$p_{i,h;j,k} = P \text{ (size} = N_j \text{ and number of A alleles} = k \text{ at generation } t+1 \mid \text{size} = N_i \text{ and number of A alleles} = h \text{ at generation } t)$$

independent of t, where $0 \leqslant h \leqslant N_i$ and $0 \leqslant k \leqslant N_j$. Progeny are obtained by binomial sampling with N_j trials, so

$$p_{i,h;j,k} = c_{ij} \binom{N_j}{k} \left[q + \frac{s_{ij}q(1-q)}{1+s_{ij}q}\right]^k \left[1 - q - \frac{s_{ij}q(1-q)}{1+s_{ij}q}\right]^{N_j-k},$$

where $q = h/N_i$, the gene frequency in state i. The $p_{i,h;j,k}$ are elements of a matrix $\mathbf{P}$, with dimensions $R \times R$, where

$$R = \sum_{i=1}^{r} (N_i + 1).$$

We wish to compute the probability of ultimate absorption of the allele A, which depends on both the size of the initial population and the initial gene frequency. We define $u_{i,h} = P$ (A is absorbed | initial population size $= N_i$ and initial number of A alleles $= h$). The vector with elements $u_{i,h}$ is denoted $\mathbf{U}$, with the same ordering of size and gene number used in P. If we let $\mathbf{U}^{(t)}$ denote the vector of mean gene frequencies at generation t, it follows that $\mathbf{U} = \lim_{t\to\infty}\mathbf{U}^{(t)}$, since the population eventually reaches complete fixation. These and other vectors or matrices are defined in Table I.

TABLE I

Definition of Matrices

Matrix	Dimension	Elements
C	$r \times r$	c_{ij} (transition probabilities for population size)
S	$r \times r$	s_{ij} (selective values)
P	$R \times R$	$p_{i,h;j,k}$ (transition probabilities)
U	$R \times 1$	$u_{i,h}$ (fixation probabilities)
V	$R \times 1$	$v_{i,h} = h/N_i$
W	$R \times 1$	$w_{i,h} = h(N_i - h)/N_i^2$
X	$R \times 1$	$x_{i,h} = h(N_i - h)(N_i - 2h)/N_i^3$
E	$R \times R$	$e_{i,h;j,k} = c_{ij}\binom{N_j}{k}(h/N_i)^k(1 - h/N_i)^{N_j-k}$
F	$R \times R$	$f_{i,h;j,k} = (s_{ij}/s)(k - N_j h/N_i)\, e_{i,h;j,k}$
G	$R \times R$	$g_{i,h;j,k} = (s_{ij}/s)^2[\frac{1}{2}k(k-1) + \frac{1}{2}N_j(N_j - 1)(h/N_i)^2 - N_j k(h/N_i) + N_j(h/N_i)^2]\, e_{i,h;j,k}$
l	$r \times 1$	$l_i = \sum_{j=1}^{r} c_{ij}(s_{ij}/s)$
m	$r \times 1$	$m_i = \sum_{j=1}^{r} c_{ij}(s_{ij}/s)^2$
A	$r \times r$	$a_{ij} = c_{ij}(1 - 1/N_j)$
B	$r \times r$	$b_{ij} = c_{ij}(1 - 1/N_j)(1 - 2/N_j)$
D	$r \times r$	$d_{ij} = (s_{ij}/s)\, c_{ij}(1 - 1/N_j)$
γ	$r \times 1$	$\gamma_j = c_{ij}$ (independent of i)
π	$r \times 1$	$\pi_j =$ stationary probabilities

In the initial generation the vector of gene frequencies is $\mathbf{V}$ (Table I); after t generations this becomes $\mathbf{U}^{(t)} = \mathbf{P}^t\mathbf{V}$ and

$$\mathbf{U} = \lim_{t\to\infty} \mathbf{P}^t\mathbf{V}. \tag{1}$$

Equation (1) represents a general method for finding the absorption probabilities. However this, or any related operation on the full matrix $\mathbf{P}$, could consume a large quantity of computer time and storage if the population can be in many size and gene-frequency states. We use a series expansion of $\mathbf{P}$ to develop an approximate result.

The elements of $\mathbf{P}$ can be expanded in a power series of the s_{ij}, in which we assume the s_{ij} are small and ignore terms higher than s_{ij}^2. We have

$$p_{i,h,j,k} = c_{ij}\binom{N_j}{k} q^k(1-q)^{N_j-k}\{1 + s_{ij}(k - N_jq) + s_{ij}^2[\tfrac{1}{2}k(k-1) - k(N_j-1)q + \tfrac{1}{2}N_j(N_j-1)\,q^2 - q(k-N_jq)]\},$$

where $q = h/N_i$. Or, written as a matrix series, we have

$$\mathbf{P} = \mathbf{E} + s\mathbf{F} + s^2\mathbf{G},$$

where the matrices $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ are defined in Table I, and include terms such as s_{ij}/s, where $s = \max(|\, s_{ij}\,|)$. Thus

$$\mathbf{U}^{(t)} = (\mathbf{E} + s\mathbf{F} + s^2\mathbf{G})\mathbf{U}^{(t-1)} \tag{2}$$

plus terms in s^3, s^4, etc.

Equation (2) is now used repeatedly, starting from $\mathbf{U}^{(0)} = \mathbf{V}$, to obtain $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$,..., and at each generation powers of s higher than s^2 are ignored. The method consists essentially of iteration on the first few moments of the gene-frequency distribution, and is given in detail in Appendix A. The final results are expressed most simply by using a special matrix operation, denoted $\bullet$. Let $\mathbf{z}$ be a column vector of dimension r, with elements z_i, and let $\mathbf{Y}$ be a vector of dimension $K = \sum_{i=1}^{r} K_i$ partitioned into r subvectors $\mathbf{Y}_i$ of dimension K_i. The operation is defined such that

$$\mathbf{z} \bullet \mathbf{Y} = \begin{pmatrix} z_1\mathbf{Y}_1 \\ z_2\mathbf{Y}_2 \\ \vdots \\ z_r\mathbf{Y}_r \end{pmatrix},$$

i.e., the result of the operation is a vector in which each block of $\mathbf{Y}$ is multiplied by the appropriate scalar element of $\mathbf{z}$. In usual matrix terms $\mathbf{z}$ would be replaced by a diagonal matrix of the same dimension as $\mathbf{Y}$ in which the first K_1 elements

are z_1, the next K_2 are z_2 and so on, with • replaced by standard matrix multiplication.

The main result is given in Eq. (A3) of Appendix A; it is

$$\mathbf{U}^{(t)} = \mathbf{V} + s(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^t)\boldsymbol{l} \bullet \mathbf{W} + \tfrac{1}{2}s^2(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B}^t)\mathbf{m} \bullet \mathbf{X}$$

$$- \tfrac{1}{2}s^2(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^t)\mathbf{m} \bullet \mathbf{W} + s^2 \sum_{i=0}^{t-2} \sum_{j=0}^{i} \mathbf{B}^j\mathbf{D}\mathbf{A}^{i-j}\boldsymbol{l} \bullet \mathbf{X} \qquad (3)$$

and from (3)

$$\mathbf{U} = \lim_{t\to\infty} \mathbf{U}^{(t)}$$

$$= \mathbf{V} + s(\mathbf{I} - \mathbf{A})^{-1}\,\boldsymbol{l} \bullet \mathbf{W} + \tfrac{1}{2}s^2(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{m} \bullet \mathbf{X} - \tfrac{1}{2}s^2(\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{m} \bullet \mathbf{W}$$
$$+ s^2(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{D}(\mathbf{I} - \mathbf{A})^{-1}\,\boldsymbol{l} \bullet \mathbf{X}, \qquad (4)$$

where terms in s^3, s^4, etc., are ignored. The method can be extended to give terms of higher order, but the algebra is very tedious.

Numerical examples in which (4) is evaluated are given later in the text.

Bounds on the values of the expressions in Eq. (4) can be obtained by a method outlined in Appendix B. We have defined $s = \max(|\, s_{ij} \,|)$, similarly let $s' = \min(|\, s_{ij} \,|)$, $N = \max(N_i)$ and $N' = \min(N_i)$ such that $s' \leqslant |\, s_{ij} \,| \leqslant s$, $N' \leqslant N_i \leqslant N$, $i, j = 1,\ldots, r$.

From Appendix B

$$N's'\mathbf{1} \leqslant |\, s(\mathbf{I} - \mathbf{A})^{-1}\,\boldsymbol{l} \,| \leqslant Ns\mathbf{1},$$

$$\tfrac{1}{3}N'(s')^2(1 - 2/3N')^{-1}\,\mathbf{1} \leqslant |\, s^2(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{m} \,| \leqslant \tfrac{1}{3}Ns^2(1 - 2/3N)^{-1}\,\mathbf{1},$$

$$N'(s')^2\mathbf{1} \leqslant |\, s^2(\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{m} \,| \leqslant Ns^2\mathbf{1}, \qquad (5)$$

$$\tfrac{1}{3}(N's')^2(1 - 1/N')(1 - 2/3N')^{-1}\,\mathbf{1} \leqslant |\, s^2(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{D}(\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{1} \,|$$
$$\leqslant \tfrac{1}{3}(Ns)^2(1 - 1/N)(1 - 2/3N)^{-1}\,\mathbf{1},$$

where $\mathbf{1}$ is a vector with all elements unity, and the inequalities in (5) apply to all elements of the vectors. If population sizes are large but s' and s are of similar order to $1/N$ and $1/N'$, terms of order Ns^2 may be ignored relative to those in Ns or N^2s^2. Using (5) we can reduce (4) to

$$\mathbf{U} = \mathbf{V} + s(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{l} \bullet \mathbf{W} + s^2(\mathbf{I} - \mathbf{B})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{l} \bullet \mathbf{X}. \qquad (6)$$

The accuracy of the approximations given by (3), (4) or (6) is difficult to determine, but we can make some useful inferences by an indirect approach. Consider the case where the population has a constant size (N) and selective value (s). We can now define the probability of absorption as $u(q)$ for a gene of

initial frequency q; in our matrix format $q = v_{ih} = h/N$ and $u(q) = u_{ih}$. Equation (4) then reduces to the first three terms of an expression derived by Narain and Robertson (1969)

$$u(q) = q + Ns(1 - s/2)\,q(1 - q) + (Ns)^2\,[(2N - 1)/(6N - 4)]q\,(1 - q)(1 - 2q) \tag{7}$$

plus terms containing s^3, s^4, etc., and we note that the coefficients of q in (7) are the same as those found in the bounds (5) for the matrix expressions. Now the probability of fixation for a population of constant size is given by Kimura (1957 and 1962) from the diffusion equation as

$$u(q) = (1 - e^{-2Nsq})/(1 - e^{-2Ns}) \tag{8}$$

that can be expanded in a power series in Ns as

$$u(q) = q + Nsq(1 - q) + \tfrac{1}{3}(Ns)^2q(1 - q)(1 - 2q) + \tfrac{1}{3}(Ns)^3q^2(1 - q)^2 + \cdots \tag{9}$$

which converges for $Ns < \pi$. Equation (7) also gives the first terms in (9) as $N \to \infty$ and Ns remains constant, and we infer that, for small s and large N, the terms excluded from (7) are given by the expansion of (9), and this has been shown for the term in $(Ns)^3$ by Narain and Robertson (1969). Thus the terms in s^3, s^4, etc., excluded from (7) are of order $(Ns)^3$, $(Ns)^4$, etc., and so (7) is a useful approximation if Ns is sufficiently small. Numerical evaluation suggests that $Ns < 1$ is required. The same requirement has to be made for our matrix results (4) or (6) with populations of variable size; but from numerical study, it does not seem necessary that the product of the *largest* population size and the *largest* selective value should be less than unity.

Successive Population Sizes Independent

The general results also simplify under less restricted conditions than constant population size and selective value. If the population sizes are independent in successive generations we may let $c_{ij} = \gamma_j$, $i, j = 1,\ldots, r$. Then $a_{ij} = \gamma_j(1 - 1/N_j)$ and it can be shown that

$$\mathbf{A}^t = (1 - 1/N^*)^{t-1}\mathbf{A} \quad \text{and} \quad (\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + N^*\mathbf{A} \tag{10}$$

where $N^* = (\sum_j \gamma_j/N_j)^{-1}$, the harmonic mean of population size. Similarly $(\mathbf{I} - \mathbf{B})^{-1} = \mathbf{I} + \tfrac{1}{3}N^*(1 - 2N^*/3N^{**})^{-1}\mathbf{B}$, where $N^{**} = (\sum_j \gamma_j/N_j^2)^{-1}$. These results can be substituted into (3), (4), or (6) and the mean frequencies or absorption probabilities easily evaluated.

With certain assumptions further simplification is possible. By arrangement of (10), we find that the i-th element of the vector $s(\mathbf{I} - \mathbf{A})^{-1}I$ is

$$N^*\bar{s} + \bar{s}_i - \sum_j \gamma_j \bar{s}_j N_j/N^*, \tag{11}$$

where $\bar{s}_i = \sum_j \gamma_j s_{ij}$ and $\bar{s} = \sum_i \gamma_i \bar{s}_i$. If the coefficients of variation of the N_i and s_i are small, and the N_i sufficiently large that terms in $1/N_i^2$ can be ignored relative to $1/N_i$, then (11) is well approximated by $N^*\bar{s}$, and (6), in scalar notation, reduces to

$$u(q) = q + N^*\bar{s}q(1 - q) + \tfrac{1}{3}(N^*\bar{s})^2 q(1 - q)(1 - 2q). \tag{12}$$

We obtain this result by a different argument in the next section.

Diffusion Approximation

Kimura (1962) has used a continuous approximation based on the diffusion equation to find absorption probabilities in populations of constant size, but with variable selection coefficient, and Chia (1968) has used the diffusion equation for populations with cyclical changes in size and selection coefficient. With certain restrictions we can use Kimura's methods for the general model described herein.

If all the population size states (N_i) are transient and can be reached from any other state, the probability that the population has a specific size reaches a stationary value, independent of generation (Feller, 1957). Letting

$$\pi_j = \lim_{t \to \infty} P \text{ (population size } = N_j), \qquad j = 1,..., r,$$

the π_j satisfy the following equations, since they are eigenvectors of $\mathbf{C}$,

$$\pi_j = \sum_{i=1}^{r} \pi_i c_{ij}, \qquad j = 1,..., r,$$

$$\sum_{j=1}^{r} \pi_j = 1,$$

(when $c_{ij} = \gamma_j$, independent of i, then $\pi_j = \gamma_j$). In the stationary state the mean selective value is $\bar{s} = \sum_i \pi_i \sum_j c_{ij} s_{ij}$, and the harmonic mean of population size is $N^* = (\sum_i \pi_i / N_i)^{-1}$. If the N_i are all of the same order of magnitude (N) and N is large, and if the s_i are of order $1/N$, the mean, $M_{\delta q}$ and variance, $V_{\delta q}$, of change in gene frequency in the stationary state become

$$M_{\delta q} = \bar{s}q(1 - q) + 0(1/N^2),$$

$$V_{\delta q} = h^2\bar{s}^2q^2(1 - q)^2 + q(1 - q)/N^* + 0(1/N^2),$$

where h is the coefficient of variation of selective value (modified from Kimura, 1962). If h is of order smaller than $N^{1/2}$, e.g., approximately unity,

$$V_{\delta q} = q(1 - q)/N^* + 0(1/N^2).$$

When the number of generations taken to reach the stationary state is small relative to N, the initial size of the population will have negligible effect on the absorption probabilities. Inserting the values of $M_{\delta q}$ and $V_{\delta q}$ we have obtained into Kimura's (1962) general formula, we obtain

$$u(q) = (1 - e^{-2N^*\bar{s}q})/(1 - e^{-2N^*\bar{s}}). \tag{13}$$

In this result the effective size (N_e) of Kimura is simply the harmonic mean of the distribution of population size, and the selective value is the arithmetic mean of its distribution. We notice that (12) forms the first three terms of an expansion of (13) in powers of Ns, although the assumptions differ slightly. Equation (13) can be expected to hold even for large values of the product $N^*\bar{s}$; and it therefore complements Eqs. (4) or (6) which are useful only when $N^*\bar{s}$ is small. It includes the results of Ewens (1967), Chia (1968) and Kimura (1970) for populations with cyclical changes in size.

Example: Normally Distributed Fitnesses

Let us now consider a simple, and perhaps meaningful biological model, in which there is an association between selective value and changes in population size. We shall use this to illustrate the methods derived in the previous sections of the paper.

Imagine that each individual in the population leaves a constant number, k, progeny, independent of population size. The proportion of these which survive to maturity and therefore the population size in the next generation is influenced by the prevailing climatic and nutritional conditions, so that a proportion of $N_j/\beta N_i$ survive when the population changes in size from N_i to N_j. We use an argument proposed by Haldane (1931) and commonly used in the theory of selection for quantitative traits to define the selective value in terms of the mean proportion surviving. Assume that individuals of genotype a have a "fitness" variable with an $N(\mu, \sigma^2)$ distribution and genotypes A an $N(\mu + \alpha\sigma, \sigma^2)$. The climatic conditions determine the minimum "fitness" which an individual needs in order to live. If α is small, such that terms of $0(\alpha^2)$ can be ignored relative to α, it can be shown that

$$s_{ij} = \alpha z_{ij}/p_{ij}, \tag{14}$$

where $p_{ij} = N_j/\beta N_i$ is the proportion surviving; and z_{ij} is the corresponding ordinate of the standardized normal density.

For a numerical example assume that two population sizes are possible, $N_1 = 50$ and $N_2 = 100$, that the fertility coefficient is $\beta = 10$ and $\alpha = 0.001$. The transition probability matrices of size and selective value are taken as

$$\mathbf{C} = \begin{pmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{pmatrix} \quad \text{and} \quad \mathbf{S} = 10^{-3} \times \begin{pmatrix} 1.755 & 1.400 \\ 2.063 & 1.755 \end{pmatrix},$$

therefore

$$\mathbf{A} = \begin{pmatrix} 0.686 & 0.297 \\ 0.098 & 0.891 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 0.65856 & 0.29106 \\ 0.09408 & 0.87318 \end{pmatrix}.$$

Letting $s = 0.001$ (for convenience, since s need not be the largest s_{ij} in computations) we have

$$l = \begin{pmatrix} 1.648 \\ 1.786 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} 2.744 \\ 3.198 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 1.2039 & 0.4158 \\ 0.2021 & 1.5638 \end{pmatrix}.$$

Substituting into (4)

$$\mathbf{U} = \mathbf{V} + 10^{-3} \begin{pmatrix} 138.7 \\ 141.1 \end{pmatrix} \bullet \mathbf{W} + 10^{-6} \begin{pmatrix} 40 \\ 42 \end{pmatrix} \bullet \mathbf{X} - 10^{-6} \begin{pmatrix} 122 \\ 124 \end{pmatrix} \bullet \mathbf{W} + 10^{-6} \begin{pmatrix} 6344 \\ 6667 \end{pmatrix} \bullet \mathbf{X}. \tag{14}$$

Thus, for a population with initial size 50 and gene frequency q, Eq. (4) gives

$$u(q) = q + 0.1386\, q(1 - q) + 0.0064\, q(1 - q)(1 - 2q),$$
$$u(0.2) = 0.22278,$$

and for an initial size of 100, $u(0.2) = 0.22320$. Taking only the terms in (6), $u(0.2) = 0.22280$ and 0.22321 for $N = 50$ and 100, respectively.

In order to evaluate expressions (12) and (13) for the same example we obtain: $\pi' = (0.25\ 0.75)$, $\bar{s} = 1.751 \times 10^{-3}$, $N^* = 80$ and $N^*\bar{s} = 0.1401$, and hence $u(0.2) = 0.22312$ from (12) and 0.22309 from (13). For this example we have found that all the approximations agree well with each other, as we would expect, since N values are relatively large, Ns values small, and there is a high probability of passage from the initial size state in a few generations. The exact results have not been computed since a 152×152 matrix would be required.

A few more examples are given in Table II in which population sizes of 5 and 10 are also used, since the exact absorption probabilities can then be calculated directly by Eq. (1). In the table we note that the approximation (4) is still useful for $N^*\bar{s}$ of about 0.5, but not when as large as 2. We also see that when transitions

TABLE II

Absorption Probabilities Calculated by Different Methods for Genes of Initial Frequency 0.2 With Gene Effect α, Normally Distributed Fitnesses and Fertility Coefficient $\beta = 10$. Two Population Sizes (N_1 and N_2) Are Possible, With Harmonic Mean N^*. The Mean Selective Value is $\bar{s}$ and **C** Specifies Transition Probabilities Between Sizes.

Method (equation)					Exact (1)		Approx. (4)		Diffusion (13)
Initial size					N_1	N_2	N_1	N_2	N_1 or N_2
N_1	N_2	**C**	α	$N^*\bar{s}$					
5	10	$\begin{pmatrix}0.7 & 0.3\\0.1 & 0.9\end{pmatrix}$	0.01	0.1401	0.22084	0.22418	0.22087	0.22422	0.22309
			0.04	0.5605	0.28577	0.30094	0.28766	0.30331	0.29796
			0.16	2.2419	0.52002	0.58584	0.61726	0.71637	0.59887
50	100		0.001	0.1401	—	—	0.22278	0.22320	0.22309
5	10	$\begin{pmatrix}0.97 & 0.03\\0.01 & 0.99\end{pmatrix}$	0.01	0.1401	0.21587	0.22761	0.21589	0.22766	0.22309
50	100		0.001	0.1401	—	—	0.22057	0.22477	0.22309

between population size states are rarer, the absorption probabilities are more dependent on the initial size so the diffusion approximation is less satisfactory.

In situations where the diffusion approximation is valid, effects of variable size can be studied easily with a model of normal fitnesses. Assume that the N_i have a symmetric distribution (e.g. normal) with mean μ and variance $c^2\mu^2$, and that the correlation of sizes in successive generations is ρ. If the coefficient of variation is sufficiently small that powers of c higher than c^2 can be ignored, then $N^* = \mu(1 - c^2)$. To evaluate $\bar{s}$, we replace z_{ij}/p_{ij} in (14) by an approximate formula for the selection differential given by Smith (1969) and obtain

$$s_{ij} = \alpha[0.8 + 0.41 \ln(\beta N_i/N_j - 1)]$$

and

$$\bar{s} = \alpha\{0.8 + 0.41[\ln(\beta - 1) - c^2(1 - \rho)/(\beta - 1)^2]\}.$$

Unless $\beta - 1$ is small, it is clear that variation in the N_i has much more influence on the effective population size than selective value. A catastrophic decline in N in a single generation would have even more influence on effective size than mean selective value.

Summary

A method is derived for computing mean gene frequencies and absorption probabilities in populations of variable size, where the probabilities of transition

between the alternative population sizes can be specified. A haploid model is adopted, with two alleles at a single locus undergoing weak selection. The results are approximate, involving the first terms in a matrix series, in which the dimensions of the matrices are the number of alternative sizes the population can take, rather than the total number of population size and gene frequency states in the full model. A diffusion-equation approximation is also considered which leads to a very simple formula. It is valid when population sizes and selective values are large, but have small coefficients of variation, and when a stationary distribution of population sizes is reached quickly. A model of selective values assuming a normal distribution of fitness is used to illustrate the results.

APPENDIX A: Derivation of Approximation for $\mathbf{U}^{(t)}$

We partition the matrices $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ into block matrices $c_{ij}\mathbf{E}_{ij}$, $c_{ij}\mathbf{F}_{ij}$, and $c_{ij}\mathbf{G}_{ij}$, respectively, of dimensions $(N_i + 1) \times (N_j + 1)$, $i, j = 1,..., r$. Similarly we partition the vectors $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{X}$ (see Table I) into subvectors $\mathbf{V}_i$, $\mathbf{W}_i$, and $\mathbf{X}_i$ of dimension $N_i + 1$. Thus

$$\mathbf{E} = \begin{pmatrix} c_{11}\mathbf{E}_{11} & c_{12}\mathbf{E}_{12} & \cdots & c_{1r}\mathbf{E}_{1r} \\ \vdots & \vdots & & \vdots \\ c_{r1}\mathbf{E}_{r1} & c_{r2}\mathbf{E}_{r2} & \cdots & c_{rr}\mathbf{E}_{rr} \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_r \end{pmatrix}.$$

Setting $\mathbf{U}^{(0)} = \mathbf{V}$, we have from (2) in the text

$$\mathbf{U}^{(1)} = \mathbf{EV} + s\mathbf{FV} + s^2\mathbf{GV}, \tag{A1}$$

where terms in s^3, s^4, etc., are being ignored. Consider the matrix product $\mathbf{EV}$. The i-th element of $\mathbf{E}_{ij}\mathbf{V}_j$ is given by

$$\sum_{k=0}^{N_j} \binom{N_j}{k} \left(\frac{h}{N_i}\right)^k \left(1 - \frac{h}{N_i}\right)^{N_j - k} \left(\frac{k}{N_i}\right) = \frac{h}{N_i}.$$

Therefore $\mathbf{E}_{ij}\mathbf{V}_j = \mathbf{V}_i$,

$$\sum_{j=1}^{r} c_{ij}\mathbf{E}_{ij}\mathbf{V}_j = \mathbf{V}_i,$$

and $\mathbf{EV} = \mathbf{V}$.

Similarly $\mathbf{F}_{ij}\mathbf{V}_j = (s_{ij}/s)\mathbf{W}_i$, $\mathbf{G}_{ij}\mathbf{V}_j = \frac{1}{2}(s_{ij}/s)^2(\mathbf{X}_i - \mathbf{W}_i)$

so $$\mathbf{FV} = \mathbf{l} \bullet \mathbf{W}, \mathbf{GV} = \tfrac{1}{2}\,\mathbf{m} \bullet \mathbf{X} - \tfrac{1}{2}\,\mathbf{m} \bullet \mathbf{W},$$

where the operation $\bullet$ is defined in the text.

Substituting in (3) we obtain,

$$\mathbf{U}^{(1)} = \mathbf{V} + s\boldsymbol{l} \bullet \mathbf{W} + \tfrac{1}{2} s^2 \mathbf{m} \bullet (\mathbf{X} - \mathbf{W}).$$

We now use Eq. (2) to obtain $\mathbf{U}^{(t)}$ in later generations. It can be shown that

$$\mathbf{E}_{ij}\mathbf{W}_j = (1 - 1/N_j)\mathbf{W}_i$$

so that

$$\sum_{j=1}^{r} (c_{ij}\mathbf{E}_{ij}\boldsymbol{l}_j\mathbf{W}_j) = \sum_{j=1}^{r} (a_{ij}\boldsymbol{l}_j)\,\mathbf{W}_i$$

or

$$\mathbf{E}(\boldsymbol{l} \bullet \mathbf{W}) = \mathbf{A}\boldsymbol{l} \bullet \mathbf{W},$$

where $\mathbf{A}$ and other matrices required are defined in Table I. Further

$$\mathbf{E}^t(\boldsymbol{l} \bullet \mathbf{W}) = \mathbf{A}^t\boldsymbol{l} \bullet \mathbf{W},$$

$$\mathbf{E}^t(\mathbf{m} \bullet \mathbf{X}) = \mathbf{B}^t\mathbf{m} \bullet \mathbf{X},$$

and

$$\mathbf{F}(\mathbf{A}^t\boldsymbol{l} \bullet \mathbf{W}) = \mathbf{D}\mathbf{A}^t\boldsymbol{l} \bullet \mathbf{X}.$$

Using these and similar relationships and ignoring terms in s^3, s^4, etc., we obtain

$$\mathbf{U}^{(2)} = \mathbf{V} + s(\mathbf{I} + \mathbf{A})\boldsymbol{l} \bullet \mathbf{W} + \tfrac{1}{2} s^2[(\mathbf{I} + \mathbf{B})\mathbf{m} \bullet \mathbf{X} - (\mathbf{I} + \mathbf{A})\mathbf{m} \bullet \mathbf{W} + 2\mathbf{D}\boldsymbol{l} \bullet \mathbf{X}],$$

$$\begin{aligned}\mathbf{U}^{(t)} = \mathbf{V} &+ s(\mathbf{I} + \cdots + \mathbf{A}^{t-1})\boldsymbol{l} \bullet \mathbf{W} + \\ &\tfrac{1}{2} s^2(\mathbf{I} + \cdots + \mathbf{B}^{t-1})\mathbf{m} \bullet \mathbf{X} - \tfrac{1}{2} s^2(\mathbf{I} + \cdots + \mathbf{A}^{t-1})\mathbf{m} \bullet \mathbf{W} \\ &+ s^2[(\mathbf{I} + \cdots + \mathbf{B}^{t-2})\mathbf{D} + (\mathbf{I} + \cdots + \mathbf{B}^{t-3})\mathbf{D}\mathbf{A} + \cdots + \mathbf{D}\mathbf{A}^{t-2}]\boldsymbol{l} \bullet \mathbf{X}.\end{aligned} \tag{A2}$$

All the eigenvalues of the matrices $\mathbf{A}$ and $\mathbf{B}$ have absolute value less than unity, so

$$\begin{aligned}\mathbf{U}^{(t)} = \mathbf{V} &+ s(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^t)\boldsymbol{l} \bullet \mathbf{W} + \tfrac{1}{2}s^2(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B}^t)\mathbf{m} \bullet \mathbf{X} \\ &- \tfrac{1}{2}s^2(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^t)\mathbf{m} \bullet \mathbf{W} + s^2 \sum_{i=0}^{t-2} \sum_{j=0}^{i} \mathbf{B}^j\mathbf{D}\mathbf{A}^{i-j}\boldsymbol{l} \bullet \mathbf{X}.\end{aligned} \tag{A3}$$

Special cases are considered in the main text.

APPENDIX B: Bounds on Terms in Approximation for U

Since N and s are the largest values attained by the population size and selective value,

$$a_{ij} = c_{ij}(1 - 1/N_j) \leqslant c_{ij}(1 - 1/N)$$

and

$$l_i = \sum_j c_{ij} s_{ij}/s \leqslant \sum_i c_{ij} = 1$$

for $1 \leqslant i, j \leqslant r$. Consider the term $(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{l}$ in Eq. (4). We have

$$\begin{aligned}(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{l} &= \mathbf{I}\boldsymbol{l} + \mathbf{A}\boldsymbol{l} + \mathbf{A}^2\boldsymbol{l} + \cdots \\ &\leqslant \mathbf{1} + (1 - 1/N)\mathbf{C}\mathbf{1} + (1 - 1/N)^2\mathbf{C}^2\mathbf{1} + \cdots, \qquad \text{(B1)}\end{aligned}$$

where $\mathbf{1}$ is a vector with all elements unity. But, since $\mathbf{C}$ is stochastic,

$$\mathbf{C}^t\mathbf{1} = \mathbf{1}, t \geqslant 0$$

and from (B1)

$$(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{l} \leqslant N\mathbf{1}.$$

The other results in Eq. (5) of the text follow from similar arguments.

References

Chia, A. B. 1968. Random mating in a population of finite size, *J. Appl. Probability* **5**, 21–30.

Crow, J. F. 1954. Breeding structure of populations. II. Effective Population number. *In* "Statistics and Mathematics in Biology" (O. Kempthorne, Ed.), pp. 543–556, Hafner, New York.

Ewens, W. J. 1967. The probability of survival of a new mutant in a fluctuating environment, *Heredity* **22**, 438–443.

Feller, W. 1957. "An introduction to probability theory and its applications," Vol. 1, 2nd ed., Wiley, New York.

Haldane, J. B. S. 1931. A mathematical theory of natural and artificial selection. VII. Selection intensity as a function of mortality rate, *Proc. Cambridge Phil. Soc.* **27**, 131–136.

Hill, W. G. 1970. Theory of limits to selection with line crossing. *In* "Mathematical Topics in Population Genetics" (K. Kojima, Ed.), pp. 210–245, Springer-Verlag, Heidelberg.

Karlin, S. 1968. Rates of approach to homozygosity for finite stochastic models with variable population size, *Amer. Natur.* **102**, 443–455.

Kimura, M. 1957. Some problems of stochastic processes in genetics, *Ann. Math. Statist.* **28**, 882–901.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population, *Genetics* **47**, 713–719.

Kimura, M. 1970. Stochastic processes in population genetics, with special reference to distribution of gene frequencies and probability of gene fixation. *In* "Mathematical Topics in Population Genetics" (K. Kojima, Ed.), pp. 178–209, Springer-Verlag, Heidelberg.

NARAIN, P. AND ROBERTSON, A. 1969. Limits and duration of response to selection in finite populations; the use of transition probability matrices, *Indian J. Hered.* **1**, 1–19.

SMITH, C. 1969. Optimum selection procedures in animal breeding, *Anim. Prod.* **11**, 433–442.

WRIGHT, S. 1939. Statistical genetics in relation to evolution. *In* "Actualités scientifiques et industrielles," Monogr. 802, Herman, Paris.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.

8

Design of experiments to estimate heritability by regression of offspring on selected parents

by

William G. Hill

Reprinted from
BIOMETRICS
THE BIOMETRIC SOCIETY, Vol. 26, No. 3. September 1970

286 NOTE:

Design of Experiments to Estimate Heritability by Regression of Offspring on Selected Parents

WILLIAM G. HILL

Institute of Animal Genetics, Edinburgh EH9 3JN, Scotland

SUMMARY

Experimental designs are given for estimating heritability by offspring-parent regression when parents can be selected and mated assortatively. Relative to designs in which selection is not practised, but the same total number of parents and progeny are recorded, the variance of the heritability estimate may be approximately halved by selecting only the best and poorest 10% or so of parents and using larger family sizes. Considerable departures can be made from the best design with small effects on efficiency.

The heritability of a quantitative trait may be estimated by the regression of progeny on parental performance. Since the sampling variance of the estimate of any linear regression coefficient is inversely proportional to the sum of squares for the independent variate, we might improve our heritability estimate by rearing and measuring a relatively large number of potential parents, but selecting only the best and poorest for mating. But a cost is incurred in measuring the discarded parents, so we might expect there to be an optimum intensity of selection which should be practised such that the sampling variance of the regression estimate is minimised for a given total expenditure in rearing and measuring parents and progeny. This design problem has apparently not been investigated, and is discussed here. Two related aspects have received attention: Latter and Robertson [1960] derived expressions for optimum progeny family size for offspring-parent regression when no selection is practised, and Soller and Genizi [1967] and Hill [1970] have discussed the optimum selection intensity in selection experiments in which family structure is ignored. Also Reeve [1961] has shown that selection or assortative mating should cause only a negligible bias to estimates of heritability from regression, so long as gene effects on the quantitative trait are small relative to its phenotypic standard deviation.

TABLE 2

PROBABILITY OF CORRECT CLASSIFICATION FOR THE TWO METHODS (AVERAGES AND RANKS WITH WEIGHTING COEFFICIENT $\alpha = 1$) FOR VARYING VALUES OF Δ/σ, THE STANDARDIZED DIFFERENCE IN THE MEAN LENGTHS OF THE TWO PAIRS OF CHROMOSOMES, AND n, THE NUMBER OF CELLS EXAMINED

			Δ/σ						
			0	0.2	0.5	1.0	2.0	3.0	∞
n	1	Averages	0.5	0.546	0.614	0.718	0.876	0.958	1.0
		Ranks	0.5	0.544	0.610	0.716	0.886	0.971	1.0
	5	Averages	0.5	0.602	0.741	0.902	0.995	1.000	1.0
		Ranks	0.5	0.582	0.700	0.857	0.987	1.000	1.0
	11	Averages	0.5	0.649	0.831	0.972	1.000	1.000	1.0
		Ranks	0.5	0.618	0.775	0.938	0.999	1.000	1.0
	21	Averages	0.5	0.702	0.907	0.996	1.000	1.000	1.0
		Ranks	0.5	0.659	0.849	0.982	1.000	1.000	1.0
	31	Averages	0.5	0.740	0.946	0.999	1.000	1.000	1.0
		Ranks	0.5	0.689	0.894	0.995	1.000	1.000	1.0

n. Table 2 also shows that, providing $\Delta/\sigma > 1.0$, measuring the four chromosomes in about 20 cells should lead to a false diagnosis rate of less than 4 in 1,000. The only ways of lowering the false diagnosis rate are to adopt some sequential procedure, measure more cells, or improve the experimental techniques so as to increase the value of Δ/σ.

ACKNOWLEDGEMENTS

I am grateful to Mrs. P. Cooke for discussions about the problem and for supplying the data from which values of Δ/σ were calculated, and to Mr. D. J. Pike for assistance with computation of the probabilities in Table 1.

UN PROBLEME DE CLASSIFICATION DES CHROMOSOMES HUMAINS

RESUME

Il est souvent nécessaire en génétique médicale de classer les chromosomes humains. Dans le problème abordè dans ce travail, un chromosome appartient à une parmi deux paires possibles de chromosomes. Pour effectuer le diagnostic on doit le classer dans l'une ou l'autre paire. On sait que les deux paires de chromosomes diffèrent par leur longueur. Deux méthodes de classement sont présentées et comparés. La première compare la moyenne de la longueur du chromosome particulier calculée sur un certain nombre de cellules à la longueur moyenne des 3 autres chromosomes appartenant aux deux paires. La seconde méthode utilise le rang (1, 2, 3 ou 4) de la longueur du chromosome particulier pour chaque cellule. On en déduit les probabilités de classement erroné et on les calcule pour différentes tailles de l'échantillon de cellules et diverses différences entre les longueurs des deux paires de chromosomes. La méthode utilisant la moyenne est généralement supérieure à la méthode de 'rang'. La relation des deux méthodes avec celle du maximum de vraisemblance est discutée.

For simplicity we shall assume that the cost of obtaining the heritability estimate is proportional to the total number of parents and progeny.

REGRESSION OF OFFSPRING ON MID-PARENT WITH ASSORTATIVE MATING

Imagine that M males and M females are measured initially, and from these the highest scoring pM and lowest scoring pM of each sex are selected. Thus p is the proportion selected in each direction, and assuming phenotypes are normally distributed, let x and i be the associated abscissa and selection differential on the standardised normal curve. The selected individuals are mated assortatively in pairs and n progeny are recorded from each mating. The total number of individuals measured over the two generations is then $T = 2M(1 + np)$. Let the phenotypic variance be σ^2 for each sex. The total sum of squares of scores among male or female parents is $2Mp(1 + ix)\sigma^2$ which is comprised of two parts: $2Mpi^2\sigma^2$ between the means of high and low selected groups, and $2Mp[1 - i(i - x)]\sigma^2$ within selected groups (Pearson [1903]). Strictly the sum of squares should be reduced by a small proportion since only a finite number are selected, but we shall assume the experiment is sufficiently large that this can be ignored. With perfect assortative mating the sum of squares among mid-parent values is also $2Mp(1 + ix)\sigma^2$. This sum of squares may be a slight overestimate in small experiments, where a correlation of one between scores of mates is unlikely to be achieved. For an additive trait with heritability h^2, the variance of observed family means about regression is comprised of two parts: $(\sigma^2/n)(1 - \frac{1}{2}h^2)$ within families, and $\frac{1}{2}h^2\sigma^2(1 - h^2)$ for genotypic means about regression, to give in all

$$\frac{\sigma^2}{n}[1 + \tfrac{1}{2}(n - 1)h^2 - \tfrac{1}{2}nh^4].$$

The estimate of heritability, $\hat{h}^2$, equals the observed regression coefficient, so

$$V(\hat{h}^2) = \frac{(np + 1)[1 + \frac{1}{2}(n - 1)h^2 - \frac{1}{2}nh^4]}{Tnp(1 + ix)}. \tag{1}$$

We wish to minimise $V(\hat{h}^2)$ by appropriate choice of n and p for given T. Assuming n takes continuous values, we obtain

$$\frac{\partial V(\hat{h}^2)}{\partial n} = 0: \qquad n^2p = \frac{1 - \frac{1}{2}h^2}{\frac{1}{2}h^2(1 - h^2)} \tag{2}$$

$$\frac{\partial V(\hat{h}^2)}{\partial p} = 0: \qquad n = \frac{x^2}{p(1 + ix - x^2)}. \tag{3}$$

Equations (2) and (3) can be solved simultaneously and a relative minimum found for $V(\hat{h}^2)$ at which n and p are functions of h^2. The optimum values of n together with the sampling variances of the estimates are given for a range of h^2 values in Table 1. The associated p values are given in Table 2 as a function of n. The optimum p values for given n are independent of h^2, and were obtained by trial and error using (3). Thus Table 2 can also be used

TABLE 1

OPTIMUM FAMILY SIZE, n, AND SAMPLING VARIANCE, $v = 100\ TV(h^2)$, FOR ALTERNATIVE METHODS. R = RANDOM MATING, A = ASSORTATIVE MATING WITHIN SELECTED GROUPS.

h^2		.05		.1		.2		.4		.6		.8	
Selection	mating	n	v	n	v	n	v	n	v	n	v	n	v
Mid-parent													
No	R	9	291	6/7	325	5	364	4	384	3	353	4	276
No	A	9	145	6/7	163	5	182	4	192	3	177	4	138
Yes	R	28	65	17	84	11	107	8	125	7	118	8	88
Yes	A	27	64	17	82	11	104	8	121	7	114	8	85
Single-parent													
(a) No	R	9	486	7	518	5	552	4	570	4	545	4/5	480
Yes	R	39	157	23	193	15	235	11	268	10	259	13	212
(b) No	R	6	523	4/5	570	3	624	2	672	2	672	2	648
Yes	R	22	199	13	252	8	320	5	394	4	425	3	425
(c) No	R	9	470	7	506	5	528	4	520	3	469	4	380
Yes	R	39	156	23	190	14	228	10	251	9	232	11	179

TABLE 2

OPTIMUM PROPORTION (%) TO SELECT IN PARENTAL GENERATION FOR SPECIFIED PROGENY FAMILY SIZE (n): MA REGRESSION ON MID-PARENT WITH ASSORTATIVE MATING, MR REGRESSION ON MID-PARENT WITH RANDOM MATING, S REGRESSION ON SINGLE PARENT

n	MA	MR	S	n	MA	MR	S
1	27.1	23.6	22.0	15	8.0	7.7	5.2
2	22.0	19.4	16.1	16	7.7	7.4	5.0
3	18.3	16.8	13.3	17	7.5	7.2	4.8
4	16.1	15.0	11.4	18	7.2	6.9	4.6
5	14.5	13.6	10.1	19	7.0	6.7	4.5
6	13.3	12.5	9.1	20	6.7	6.5	4.3
7	12.3	11.6	8.4	22	6.4	6.2	4.0
8	11.4	10.8	7.7	24	6.0	5.8	3.8
9	10.7	10.2	7.2	26	5.7	5.5	3.6
10	10.1	9.7	6.7	28	5.5	5.3	3.4
11	9.6	9.2	6.4	30	5.2	5.1	3.3
12	9.1	8.7	6.0	32	5.0	4.9	3.1
13	8.7	8.4	5.7	36	4.6	4.5	2.9
14	8.4	8.0	5.5	40	4.3	4.2	2.7

to design experiments in which the optimum family sizes cannot be achieved. For example, if the heritability is thought to be about 0.2, the best design requires selection of the highest and lowest scoring 9.6% of parents and taking families of size 11, when $V(\hat{h}^2) = 1.04/T$. However, if the maximum family size which can be attained is 4, then 16.1% of parents should be chosen.

Table 1 also includes values for the optimum family sizes (Latter and Robertson [1960]) and sampling variance of the estimates for designs in which parents are not selected ($p = 0.5$). Selection among the parents is seen to increase efficiency by a factor of about 2 with intermediate heritabilities, and requires family sizes 2 – 3 times as large.

In practice we may have a very poor prediction of h^2, or the best design may not be feasible because of restrictions imposed by the facilities or by the fertility of the species. In Figure 1, $V(\hat{h}^2)$ for several designs are compared with the optimum design for each h^2 value. For example, we see that $n = 10$, $p = 10\%$ is very efficient over a wide range of heritability values, and this design has another feature which may be useful, in that the same number

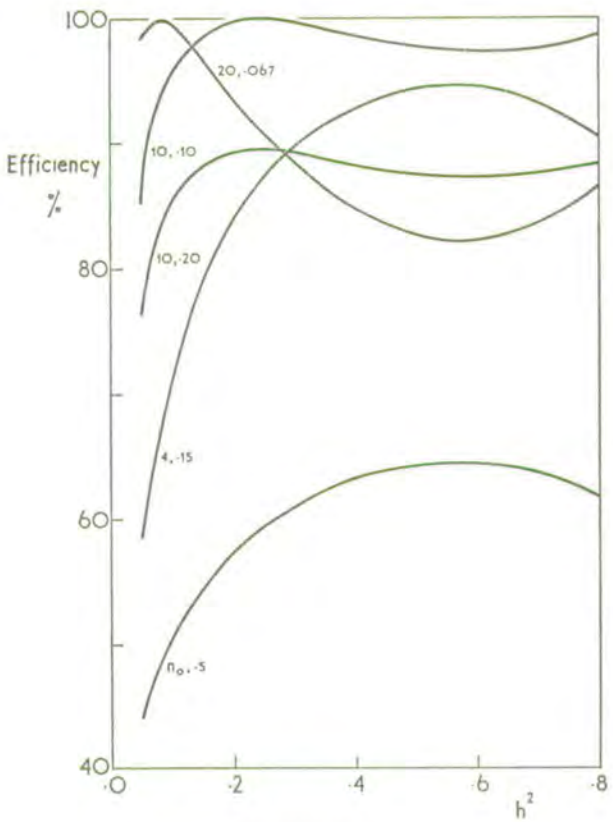

FIGURE 1

EFFICIENCY OF ALTERNATIVE DESIGNS (n, p) RELATIVE TO THE OPTIMUM DESIGN, EACH WITH REGRESSION ON MID-PARENT AND ASSORTATIVE MATING. WHERE PARENTS ARE UNSELECTED (n_0, 0.5) THE BEST APPROPRIATE FAMILY SIZE IS USED.

of parents and progeny are measured. In general, we find that considerable departure can be made from the optimum design yet have little effect on the variance of the heritability estimate. An example is also given in Figure 1 of two designs with the same value of n, but different p. As we would expect from equation (1), their relative efficiency is independent of h^2.

REGRESSION OF OFFSPRING ON MID-PARENT WITH RANDOM MATING

Selection may be practised among the parents, yet these be mated at random within high and low selected groups. Although mating is still assortative *between* the groups, we shall restrict use of the term 'assortative' to describe the mating *within* selected groups. With random mating the sum of squares of mid-parent values within the high and low selected groups is halved, and

$$V(\hat{h}^2) = \frac{2(np + 1)[1 + \frac{1}{2}(n - 1)h^2 - \frac{1}{2}nh^4]}{Tnp(1 + ix + i^2)}.$$

Equation (2) is unchanged, and equation (3) becomes

$$\frac{\partial V(\hat{h}^2)}{\partial p} = 0: \qquad n = \frac{x^2 + 2ix - i^2}{p[1 - x^2 - ix + 2i^2]}.$$

The optimum values of p for specified n are listed in Table 2, and the best designs and their associated sampling variances are given in Table 1. With unselected parents assortative mating doubles the efficiency at no cost, but at the desired selection intensities of the order of 10% there is little further gain from assortative mating because the variance within the groups of selected parents is small.

REGRESSION OF OFFSPRING ON SINGLE PARENT

When only one parent is measured the total number of recorded individuals in the two generations is $T = M(2n\ p + 1)$, and assortative mating cannot be practised. The optimum value of p for given family size is given by

$$\frac{\partial V(\hat{h}^2)}{\partial p} = 0: \qquad n = \frac{x^2}{2p(1 + ix - x^2)}. \tag{4}$$

The solutions are summarised in Table 2.

Latter and Robertson [1960] described three situations in which heritability may be estimated by doubling the regression of offspring on single parent. These are listed below, together with $V(\hat{h}^2)$ and the equation obtained from $\partial V(\hat{h}^2)/\partial n = 0$ which is used with (4) to find the optimum design.

(a) Offspring families are half sibs and related only through the measured parent (e.g. son on sire regression in cattle).

$$V(\hat{h}^2) = \frac{2(2np + 1)[1 + \frac{1}{4}(n - 1)h^2 - \frac{1}{4}nh^4]}{Tnp(1 + ix)}$$

$$n^2p = (1 - \tfrac{1}{4}h^2)/\tfrac{1}{2}h^2(1 - h^2).$$

(b) Each measured parent has only one mate, so that the offspring are full sibs.

$$V(\hat{h}^2) = \frac{2(2np + 1)[1 + \frac{1}{2}(n - 1)h^2 - \frac{1}{4}nh^4]}{Tnp(1 + ix)}$$

$$n^2p = 1/h^2.$$

(c) All measured parents have the same mate (e.g. intra-sire daughter dam regression).

$$V(\hat{h}^2) = \frac{2(2np + 1)[1 + \frac{1}{4}(n - 2)h^2 - \frac{1}{4}nh^4]}{Tnp(1 + ix)}$$

$$n^2p = (1 - \tfrac{1}{2}h^2)/\tfrac{1}{2}h^2(1 - h^2).$$

The optimum family sizes and the sampling variances of the estimates are given in Table 1. We see that rather larger families should be used and more intense selection should be practised than when regression is on mid-parent. By setting $p = 0.5$ in the above equations we obtain the formulae of Latter and Robertson [1960] for the optimum family size with unselected parents, and the designs are summarised in Table 1. As we found with regression on mid-parent, selection of the single parent approximately halves the sampling variance but requires family sizes 2–4 times as large.

PLAN D'EXPERIENCE POUR ESTIMER L'HERITABILITE PAR LA REGRESSION DE LA DESCENDANCE SUR DES PARENTS SELECTIONNES.

RESUME

On donne des plans expérimentaux pour estimer l'héritabilité par la régression enfants-parent quand on peut sélectionner les parents et les croiser de manière assortative. Relativement aux plans dans lesquels on ne pratique pas de sélection, mais comportant le même nombre total de parents et d'enfants, on peut réduire approximativement de moitié la variance de l'héritabilité en sélectionnant seulement 10% (ou un pourcentage voisin) des meilleurs et des moins bons des parents et en utilisant des familles plus grandes. On peut faire des écarts considérables au meilleur plan en n'entrainant que des effets petits sur l'efficacité.

REFERENCES

Hill, W. G. [1970]. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics* (submitted).

Latter, B. D. H. and Robertson, A. [1960]. Experimental design in the estimation of heritability by regression methods. *Biometrics 16*, 348–53.

Pearson, K. [1903]. On the influence of natural selection on the variability and correlation of organs. *Phil. Trans. A200, 1–66.*

Reeve, E. C. R. [1961]. A note on non-random mating in progeny tests. *Genet. Res. 2*, 195–203.

Soller, M. and Genizi, A. [1967]. Optimum experimental designs for realised heritability estimates. *Biometrics 23*, 361–5.

Received March 1970

9

Estimation of heritability by both regression of offspring on parent and intra-class correlation of sibs in one experiment

by

William G. Hill and Frank W. Nicholas

BIOMETRICS *30*, 447–468
SEPTEMBER 1974

ESTIMATION OF HERITABILITY BY BOTH REGRESSION OF OFFSPRING ON PARENT AND INTRA-CLASS CORRELATION OF SIBS IN ONE EXPERIMENT

W. G. HILL AND F. W. NICHOLAS[1]

Institute of Animal Genetics, Edinburgh EH9 3JN, Scotland

SUMMARY

The analysis and design of experiments to estimate heritability when data are available on both parents and offspring are discussed. It is shown that there is a substantial positive sampling correlation between the regression of offspring on mid-parent and the covariance of full sibs estimated from the same data, and that in a hierarchical structure the covariance of half sibs has a negative correlation with the regression of offspring on dam and a positive correlation with the regression of offspring on sire.

The efficiency of alternative estimators of heritability by regression and sib covariance, pooled estimators based on these and maximum likelihood (ML) are compared. The ML estimator does not reduce the variance substantially below that from the pooled estimators, but both are often much better than either regression or sib covariance estimators alone.

The optimum designs of experiments for ML estimation are obtained. It is found that these do not differ very much from those appropriate for either offspring on parent regression or half sib covariance estimators, and that optimum designs are fairly robust against changes in parameter assumptions.

1. INTRODUCTION

In laboratory or field experiments data are sometimes available on the performance of both the parents and several of their progeny. It is then possible to estimate heritability in two ways, either from the regression of progeny on parent performance or from the intra-class correlation of sibs in the progeny generation (e.g. Falconer [1960]). In the regression method, no use is made of the variance between members of the same family, or, directly, of the variance between family means. In the intra-class correlation method, no use is made of parental performance. When all the information is available heritability is customarily estimated by both methods from the same data, but no attempt is made to find the correlation between the estimates, or to pool them to obtain a single, best estimate. Alternative estimates of heritability from the same data have been obtained by Sheridan *et al.* [1968], who commented on the poor agreement obtained between the offspring-parent and sib covariance estimates, but thought this due to sampling. Clayton *et al.* [1957] obtained the different kinds of estimates, but each from a different set of data. Alternatively all the information could be utilized to form a ML estimate, which is not commonly done in practice, but has been suggested in this context by Dr. J. Felsenstein (personal communication).

In this paper we derive formulae for the expected values of the sampling correlation between regression and intra-class correlation heritability estimates, of the variance of pooled estimates derived from these, and of ML estimates. Thus we envisage, in concept,

[1] Present address: Department of Animal Husbandry, University of Sydney, Sydney, N. S. W. 2006, Australia.

a large number of separate experiments, of identical design, in each of which a heritability is estimated by offspring-parent regression and by the covariance of sibs. The sampling correlation we compute is that between the pairs of estimates obtained in each experiment taken over the population of replicated experiments.

In section 2 we discuss the concepts and derive in some detail the formulae for a very simple situation, full sib families from pair matings. In section 3 we give without details of derivation equivalent formulae for the more involved, but more important hierarchical design in which males are each mated to several females, to give both full-sib and half-sib family groups. In section 4 we compare the efficiency of alternative estimators and in section 5 we discuss the optimum designs for estimating heritability using all the available information by ML.

We assume that random mating is practiced. For simplicity, balanced designs are considered which, though rarely encountered in field data, illustrate the principles more clearly.

2. FULL SIB STRUCTURE

If the correlation of full sibs is to be an unbiased estimator of heritability we need to assume that gene action is additive and that there is no covariance among sibs produced by common environmental (maternal) effects; and for the regression of offspring on parent to be an unbiased estimator, there must be no environmental covariance of maternal and progeny performance. We make all these assumptions here, but relax some of them in the half sib analysis discussed subsequently.

Let us assume that s pair matings are made, and that n progeny are reared from each mating. Although some information is contained in the variance between individual parents, we shall ignore this, and utilize only the parental means, X_i , $i = 1, \cdots, s$. Let Z_{ij} be the score of the jth individual in the ith family, with $j = 1, \cdots, n$. We assume that the X_i and Z_{ij} are multivariate normally distributed, each with mean μ, and that individual observations have variance σ^2. The typical variance-covariance structure, based on formulae given by Falconer [1960], is shown below:

$$\begin{array}{c} \\ X_i \\ Z_{ij} \\ Z_{ij'} \end{array} \begin{array}{c} \begin{array}{ccccc} X_i & Z_{ij} & Z_{ij'} & X_{i'} & Z_{i'j} \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2}H & \frac{1}{2}H & 0 & 0 \\ \frac{1}{2}H & 1 & \frac{1}{2}H & 0 & 0 \\ \frac{1}{2}H & \frac{1}{2}H & 1 & 0 & 0 \end{bmatrix} \end{array} \sigma^2 \qquad (1)$$

where $i \neq i'$, $j \neq j'$, and H is the heritability (h^2).

Regression and intra-class correlation

In the usual offspring-parent and sib covariance analyses the following mean squares or products are computed:

$$M_{XX} = \sum_i (X_i - \bar{X}_.)^2/(s-1), \qquad M_{XZ} = \sum_i (X_i - \bar{X}_.)(\bar{Z}_{i.} - \bar{Z}_{..})/(s-1),$$

$$M_{BZ} = n \sum_i (\bar{Z}_{i.} - \bar{Z}_{..})^2/(s-1), \qquad M_{WZ} = \sum_i \sum_j (Z_{ij} - \bar{Z}_{i.})^2/s(n-1);$$

and the following estimators of heritability may be used:

regression of offspring on mid-parent: $H_{bf} = M_{XZ}/M_{XX}$,
twice the intra-class correlation of full sibs:

$$H_{tf} = 2(M_{BZ} - M_{WZ})/[M_{BZ} + (n-1)M_{WZ}].$$

While H_{bf} is an unbiased estimator of H, H_{tf} is not, for it is the ratio of two random variables, for which only the ratio of their expectations is H. We have

$$V(\bar{Z}_{i\cdot} \mid X_i) = [\tfrac{1}{2}H(1-H) + (1 - \tfrac{1}{2}H)/n]\sigma^2,$$

and since $\sum_i (X_i - \bar{X}_{\cdot})^2/(\sigma^2/2)$ is distributed as chi-square with $s-1$ D.F.,

$$E[1/\sum_i (X_i - \bar{X}_{\cdot})^2] = 2/[(s-3)\sigma^2]$$

which can be shown directly, or inferred from Kendall and Stuart ([1973] p. 305). Hence

$$V(H_{bf}) = \frac{2 + (n-1)H - nH^2}{(s-3)n}$$

(Latter and Robertson [1960]). Here and elsewhere we shall assume that s is sufficiently large that terms of order s^{-1} can be ignored relative to 1, giving

$$V(H_{bf}) \doteq [2 + (n-1)H - nH^2]/sn. \tag{2}$$

By taking logarithms and expanding, or using Taylor's series, we obtain

$$V(H_{tf}) \doteq \frac{(2-H)^2[2 + (n-1)H]^2(sn-1)}{2s(s-1)n^2(n-1)},$$

which reduces to Fisher's ([1925] section 39) formula

$$V(H_{tf}) \doteq \frac{(2-H)^2[2 + (n-1)H]^2}{2sn(n-1)} \tag{3}$$

approximately, if s is large.

We find cov (H_{bf}, H_{tf}) by the same expansion method. For four random variables $w_1, \cdots, w_4$ with means $\mu_1, \cdots, \mu_4$ and small coefficients of variation such that terms of order $(w_i - \mu_i)^3/\mu_i^3$ and higher can be ignored, then

$$\operatorname{cov}\left(\frac{w_1}{w_2}, \frac{w_3}{w_4}\right) \doteq \frac{\mu_1\mu_3}{\mu_2\mu_4}\left[\frac{\operatorname{cov}(w_1, w_3)}{\mu_1\mu_3} - \frac{\operatorname{cov}(w_1, w_4)}{\mu_1\mu_4} - \frac{\operatorname{cov}(w_2, w_3)}{\mu_2\mu_3} + \frac{\operatorname{cov}(w_2, w_4)}{\mu_2\mu_4}\right]. \tag{4}$$

In our case we have

$$w_1 = M_{XZ},\ \mu_1 = \tfrac{1}{2}H\sigma^2; \qquad w_2 = M_{XX},\ \mu_2 = \tfrac{1}{2}\sigma^2$$

$$w_3 = 2(M_{BZ} - M_{WZ}),\ \mu_3 = nH\sigma^2; \qquad w_4 = M_{BZ} + (n-1)M_{WZ},\ \mu_4 = n\sigma^2.$$

Tallis [1959] gives a general formula for variances and covariances of mean squares and products of normal deviates. For some m_{qr}, m_{st} which are unbiased estimators of population moments with f D.F.,

$$\operatorname{cov}(m_{qr}, m_{st}) = [\operatorname{cov}(q, s)\operatorname{cov}(r, t) + \operatorname{cov}(q, t)\operatorname{cov}(r, s)]/f$$

where cov (q, s) etc. are the appropriate covariances. We have

$$\operatorname{cov}(M_{XX}, M_{BZ}) = 2n \operatorname{cov}^2(X_i, \bar{Z}_{i\cdot})/(s-1) = \tfrac{1}{2}nH^2\sigma^4/(s-1)$$

$$\operatorname{cov}(M_{XX}, M_{WZ}) = \operatorname{cov}(M_{XZ}, M_{WZ}) = 0$$

$$\operatorname{cov}(M_{XZ}, M_{BZ}) = 2n \operatorname{cov}(X_i, \bar{Z}_{i\cdot})V(\bar{Z}_{i\cdot})/(s-1) = \tfrac{1}{2}H(nH + 2 - H)\sigma^4/(s-1).$$

Substituting the above into (4), rearranging, and assuming s is large we obtain

$$\text{cov}\,(H_{bf}\,,\,H_{tf}) \doteq \frac{H(2-H)[2+(n-1)H-nH^2]}{sn} \tag{5}$$

which is, of course, approximate since high order terms are ignored in (4).

From (2) and (5) we find that, asymptotically for large s, the regression of H_{tf} on H_{bf} is given by $H(2-H)$, and from (2), (3), and (5) that the correlation between H_{tf} and H_{bf} is

$$r \doteq \frac{H\{2(n-1)[2+(n-1)H-nH^2]\}^{\frac{1}{2}}}{2+(n-1)H} \tag{6}$$

which does not depend on the number of families. With large family sizes ($n \to \infty$) and $H > 0$, equation (6) reduces to $r \doteq [2H(1-H)]^{1/2}$. In Figure 1 the correlation is shown for some values of n and H.

Some verbal but nonrigorous explanation of the positive covariance and hence correlation of the two estimators can be given. If, for example, the genetic variance among the sample of parental pairs taken exceeds its expectation $H\sigma^2/2$, then the variance between progeny means and the covariance of progeny and parental scores will both exceed their appropriate expectation, so that both H_{tf} and H_{bf} will tend to exceed H. However, both H_{tf} and H_{bf} will generally be less than H if there is reduced genetic variance among parental pairs, so there is a positive covariance between H_{tf} and H_{bf} .

It is clear from Figure 1 that the correlation between estimates of heritability from offspring on mid parent regression and from the covariance of full sibs is not trivially small unless the true heritability (H) is close to zero or, only if family sizes are very large,

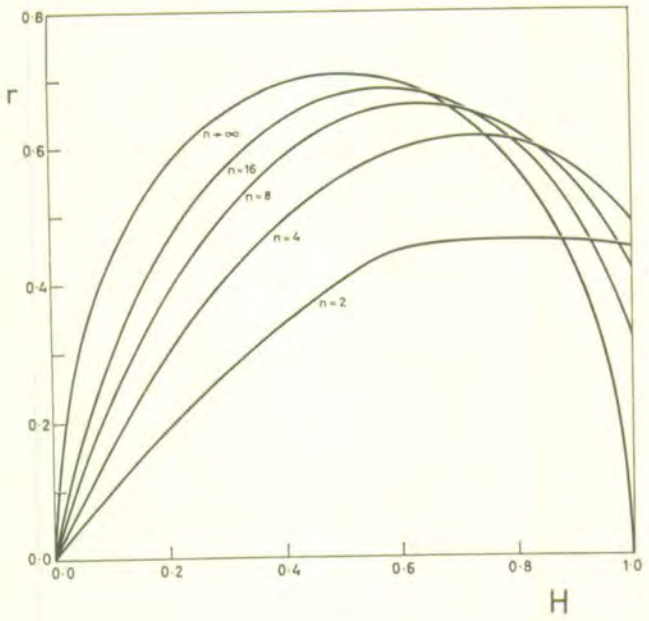

FIGURE 1

CORRELATION (r) BETWEEN ESTIMATES OF HERITABILITY USING FULL SIB FAMILIES FROM THE COVARIANCE OF FULL SIBS (H_{tf}) AND THE REGRESSION OF OFFSPRING ON PARENT (H_{bf}).

close to unity. Thus, in a single experiment in which heritability is estimated by both methods, we should expect to find a better agreement between the two estimates than if they were obtained independently. This does not imply that they have similar efficiencies as we shall see subsequently.

Maximum likelihood estimation

The available information on heritability in the experiment can be utilized by ML. We are concerned here primarily with the efficiency of such estimators, relative to using the simple regression or sib correlation estimators, rather than with the ML estimation procedure.

Let $\mathbf{V}$ of dimension $s(n+1)$ be the variance-covariance matrix of the observations, which, for simplicity in the later analysis, we take as the transformed vector:

$$(X_1, \bar{Z}_{1.}, Z_{11} - \bar{Z}_{1.}, Z_{12} - \bar{Z}_{1.}, \cdots, Z_{1,n-1} - \bar{Z}_{1.}, \cdots, X_s, \bar{Z}_{s.},$$
$$Z_{s1} - \bar{Z}_{s.}, \cdots, Z_{s,n-1} - \bar{Z}_{s.})'.$$

Since families are distributed independently, $\mathbf{V}$ is block diagonal, with the block $\mathbf{V}_i$ of dimension $n+1$ specifying the variance-covariance structure of a single family. We can write $\mathbf{V} = \mathbf{I} * \mathbf{V}_i\sigma^2$, where * denotes direct product (Searle [1966]). From the model (1)

$$\mathbf{V}_i = \begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \left(1 - \frac{H}{2}\right)\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) \end{bmatrix}, \qquad i = 1, \cdots, s$$

where $\mathbf{I}$ (the identity matrix) and $\mathbf{J}$ (with all elements unity) are of dimension $n-1$; and $\mathbf{T}$ of dimension 2 is given by

$$\mathbf{T} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}H \\ \frac{1}{2}H & \frac{1}{2}H + (1 - \frac{1}{2}H)/n \end{bmatrix}.$$

Noting that $E(X_i) = E(\bar{Z}_{i.}) = \mu$ and $E(Z_{ij} - \bar{Z}_{i.}) = 0$, the log likelihood becomes

$$\text{Log } L = -\tfrac{1}{2}s(n+1)(\log 2\pi + \log \sigma^2) - \tfrac{1}{2}s \log |\mathbf{T}| + \tfrac{1}{2}s \log n - \tfrac{1}{2}s(n-1)$$
$$\times \log (1 - \tfrac{1}{2}H) - (1/2\sigma^2) \sum_{i=1}^{s} [(\mathbf{y}_i - \mu\mathbf{1})'\mathbf{T}^{-1}(\mathbf{y}_i - \mu\mathbf{1}) + \sum_{j=1}^{n} (Z_{ij} - \bar{Z}_{i.})^2/(1 - \tfrac{1}{2}H)] \quad (7)$$

where $\mathbf{y}_i' = (X_i, \bar{Z}_{i.})$, $\mathbf{1}' = (1, 1)$. Explicit solutions for the ML estimators of μ, σ^2 and H have not been found, but with any set of data estimates can be obtained numerically. For example, Felsenstein (personal communication) has written a computer program for this specific problem. However, large sample variances can be obtained in the usual way from the inverse of the matrix of expected second partial derivatives of the likelihood with respect to the parameters.

Let $\theta_1 = \mu$, $\theta_2 = \sigma^2$ and $\theta_3 = H$, and the information matrix $\mathbf{M}$ have elements

$$m_{ij} = -E(\partial^2 \log L/\partial\theta_i\partial\theta_j), \; i, j = 1, 2, 3.$$

In differentiating (7) and taking expectations we utilize some results given by Searle [1970]. In our context these are

$$E\left\{\frac{\partial}{\partial H}[(\mathbf{y}_i - \mu\mathbf{1})'(\sigma^2\mathbf{T})^{-1}(\mathbf{y}_i - \mu\mathbf{1})]\right\} = -\text{tr}\left[\mathbf{T}^{-1}\frac{\partial\mathbf{T}}{\partial H}\right] = -\partial \log |\mathbf{T}|/\partial H,$$

where tr denotes the trace; and

$$E\left\{\frac{\partial^2}{\partial H^2}[(\mathbf{y}_i - \mu\mathbf{1})'(\sigma^2\mathbf{T})^{-1}(\mathbf{y}_i - \mu\mathbf{1})]\right\} = -\mathrm{tr}\left[\mathbf{T}^{-1}\frac{\partial^2\mathbf{T}}{\partial H^2} - 2\mathbf{T}^{-1}\frac{\partial\mathbf{T}}{\partial H}\mathbf{T}^{-1}\frac{\partial\mathbf{T}}{\partial H}\right],$$

$$= -\frac{2\partial^2}{\partial H^2}\log|\mathbf{T}| + \mathrm{tr}\left[\mathbf{T}^{-1}\frac{\partial^2\mathbf{T}}{\partial H^2}\right]$$

$$= -2\partial^2 \log|\mathbf{T}|/\partial H^2,$$

since $\partial^2\mathbf{T}/\partial H^2 = 0$. We obtain from (7)

$$m_{11} = \frac{2s[n + 2 - (n + 1)H]}{\sigma^2[2 + (n - 1)H - nH^2]}, \qquad m_{22} = \frac{s(n + 1)}{2\sigma^4}$$

$$m_{12} = m_{21} = m_{13} = m_{31} = 0$$

$$m_{23} = m_{32} = \frac{s}{2\sigma^2}\left[\frac{n - 1 - 2nH}{2 + (n - 1)H - nH^2} - \frac{n - 1}{2 - H}\right]$$

$$m_{33} = s\left\{\frac{2n[2 + (n - 1)H - nH^2] + (n - 1 - 2nH)^2}{2[2 + (n - 1)H - nH^2]^2} + \frac{n - 1}{2(2 - H)^2}\right\}.$$

The estimates of μ and H are uncorrelated, since they are the mean and a function of the variance, respectively, in a mixed model (Searle [1970]). Let $V(H_{mf})$ denote the sampling variance of the ML estimator of heritability, which is given by the (3, 3) element of $\mathbf{M}^{-1}$, i.e.

$$V(H_{mf}) = m_{22}(m_{22}m_{33} - m_{23}^2)^{-1}.$$

Relative efficiency of estimators

The variance of H_{mf} is compared with that of the simple estimators H_{bf} and H_{tf} in Figure 2. The total number of observations made for the estimates is $T = s(n + 2)$, so to enable comparisons between estimates obtained for different values of n, variances are expressed as $T \cdot V(H_{mf}) = v$, for example. Thus for any experiment with T^* individuals, the variance is v/T^*. The computed sampling variance of the ML estimator is proportional to s, and we have seen that those of H_{tf} and H_{bf} are inversely proportional to $s - 1$ and $s - 3$, respectively, and approximately to s if the number of sires is large. We therefore assume that many sires are used, and the results of Figure 2 do not depend on s.

It is also possible to obtain a pooled heritability estimate, H_{pf}, as a linear weighted function of H_{bf} and H_{tf}. We take

$$H_{pf} = \alpha H_{bf} + (1 - \alpha)H_{tf} \tag{8}$$

in which α is chosen so as to minimize $V(H_{pf})$. This value of α is

$$\alpha = [V(H_{tf}) - \mathrm{cov}\,(H_{bf}, H_{tf})]/[V(H_{bf}) + V(H_{tf}) - 2\,\mathrm{cov}\,(H_{bf}, H_{tf})], \tag{9}$$

giving

$$V(H_{pf}) = [V(H_{bf})V(H_{tf}) - \mathrm{cov}^2\,(H_{bf}, H_{tf})]/[V(H_{bf}) + V(H_{tf}) - 2\,\mathrm{cov}\,(H_{bf}, H_{tf})]. \tag{10}$$

In practice only estimates of $V(H_{bf})$, $V(H_{tf})$ and $\mathrm{cov}\,(H_{bf}, H_{tf})$ are available to insert into (9), since they depend on the parameter H. An iterative procedure has to be used in which a value, $\hat{\alpha}$, is guessed, used to estimate H_{pf} from (8), and subsequently $V(H_{bf})$ etc. These values are substituted into (9), $\hat{\alpha}$ is estimated again and the process repeated.

Values of $T \cdot V(H_{pf})$ are also shown in Figure 2. Since the best weighting factor, α, is not known, the variances given in the figure may be biased downwards. While no exact formula for this bias has been obtained, a simple argument shows that it becomes proportionately smaller as s increases, and thus is negligible in large samples. Rewriting (8) as

$$H_{pf} = H_{if} + \hat{\alpha}(H_{bf} - H_{if})$$

we see that the contribution of error of estimation of α to $V(H_{pf})$ is roughly proportional to $E(H_{bf} - H_{if})^2 V(\hat{\alpha})$. Now $E(H_{bf} - H_{if})^2$ and the variance of all of the terms on the right hand side of (9), and thus $V(\hat{\alpha})$, are proportional to $1/s$, so the product $E(H_{bf} - H_{if})^2 V(\hat{\alpha})$ is proportion to $1/s^2$ and in large samples becomes a trivial part of $V(H_{pf})$. (The same arguments can be applied to the ML estimators, which are themselves weighted estimates, with the weights inaccurately determined in small samples).

In the comparisons shown by solid lines in Figure 2 it is assumed that observations have been made on both parents and progeny, giving a variance per observation of $v = s(n + 2)V(H_{if})$, for example. But if only the intra-class correlation estimator is required the $2s$ observations on parents do not have to be made, giving $v = sn\,V(H_{if})$ is an experiment of the same design. Thus an extra broken curve, denoted H_{if}' is included for the intra-class correlation showing the variance of the estimate per progeny observation. It differs from the curve H_{if} by a constant proportion $n/(n + 2)$, (and thus by a constant difference on the logarithmic scale) which is the proportion of the observations made on progeny.

With small family sizes the ML or pooled estimator based on parents and progeny is more efficient than the intra-class correlation with data collected on progeny alone. With larger family sizes the two alternatives have a similar efficiency at low heritabilities, but the estimators using information on regression are more efficient at high heritabilities (Figure 2). Except at high heritabilities the pooled estimator, H_{pf} , is almost as efficient as the ML estimator.

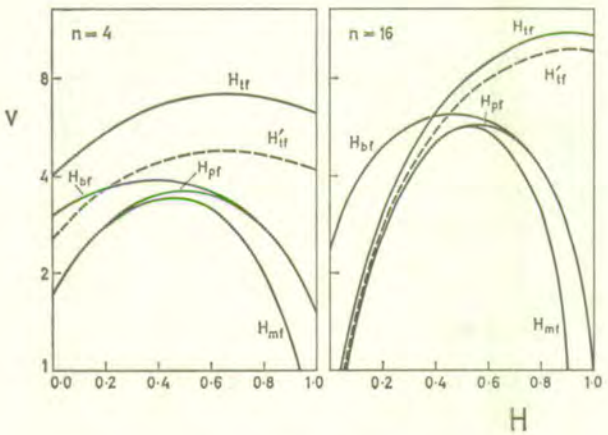

FIGURE 2

SAMPLING VARIANCES PER OBSERVATION (v) OF ALTERNATIVE HERITABILITY ESTIMATORS USING FULL SIB FAMILIES WITH DATA COLLECTED ON PARENTS AND PROGENY (SOLID CURVES); SIB COVARIANCE (H_{if}), REGRESSION ON MID-PARENT (H_{bf}), POOLED SIB COVARIANCE AND REGRESSION (H_{pf}) AND MAXIMUM LIKELIHOOD (H_{mf}). THE SAMPLING VARIANCE PER PROGENY OBSERVATION OF THE SIB COVARIANCE ESTIMATOR IS ALSO GIVEN (H_{if}').

The loss of efficiency in ML estimation from excluding the information on the individual parents can be obtained using the methods described in section 3, but omitting the environmental covariance of sibs term (K). For heritabilities near zero there is no loss in efficiency. Taking values of H of 0.1, 0.2, $\cdots$, 0.9, the greatest losses obtained were 6.5% and 7.5% for $n = 16$ and 8 respectively, both at $H = 0.7$, and 9.5%, 12.2% and 11.9% for $n = 4$, 2, and 1 respectively, all at $H = 0.9$.

3. ESTIMATORS IN A HIERARCHICAL STRUCTURE

An important assumption in our analysis of the full sib family model is that the only covariance between family members is that from additive genetic variance (i.e. $H\sigma^2/2$). Usually there is some additional covariance, $K\sigma^2$, of full sibs from two sources; maternal or other environment effects common to full sibs and nonadditive genetic effects, especially dominance (Falconer [1960]). Therefore intra-class correlation estimates of heritability are normally made from the covariance of half sibs. Regressions of progeny on parental performance do not include dominance effects, but there could be some maternal environmental covariance between progeny and dam. However, this covariance is unlikely to be of the same magnitude as the environmental covariance of sibs and we shall assume in the following analysis that it can be ignored. Thus the only major change from the simple full sib model described previously is that a term $K\sigma^2$ is added to the covariance of full sibs. We again assume there is no epistatic variance.

Let s sires each be mated to d dams with n progeny reared from each mating and we shall assume throughout that s is sufficiently large that terms in s^{-1} can be ignored relative to 1. This simplifies the formulae and makes them more directly comparable with each other. Let X_i be the measurement on sire i, Y_{ij} that on the jth dam mated to sire i, and Z_{ijk} the measurement on her kth progeny. The observations are assumed to be multivariate normally distributed with mean μ. There are no covariances between members of different sire families, and typical variances and covariances for a single family are shown below:

$$
\begin{array}{c|cccccc}
 & X_i & Y_{ij} & Z_{ijk} & Z_{ijk'} & Y_{ij'} & Z_{ij'k} \\
X_i & 1 & 0 & \frac{1}{2}H & \frac{1}{2}H & 0 & \frac{1}{2}H \\
Y_{ij} & 0 & 1 & \frac{1}{2}H & \frac{1}{2}H & 0 & 0 \\
Z_{ijk} & \frac{1}{2}H & \frac{1}{2}H & 1 & \frac{1}{2}H + K & 0 & \frac{1}{4}H \\
Z_{ijk'} & \frac{1}{2}H & \frac{1}{2}H & \frac{1}{2}H + K & 1 & 0 & \frac{1}{4}H \\
Y_{ij'} & 0 & 0 & 0 & 0 & 1 & \frac{1}{2}H \\
Z_{ij'k} & \frac{1}{2}H & 0 & \frac{1}{4}H & \frac{1}{4}H & \frac{1}{2}H & 1
\end{array} \sigma^2 \qquad (11)
$$

where $j \neq j'$, $k \neq k'$. Within this structure we shall also include the case of sex limited traits, where if no measurement is made on males, no X_i are available, or if none on females, there are no Y_{ij} . There are clearly many other relevant models which we do not consider: for example where males and females have different means and variances or where the mean performance differs between the two generations.

With this kind of data estimates of heritability can be obtained in several ways:

i) Intra-class correlation between half sibs, (The correlation between full sibs is biased.)

ii) Regression of offspring on parent performance:

   a. Progeny on dam within sires,

b. Progeny on sire,
c. Progeny on sire plus dam average,
d. Progeny on mid-parent,
e. Various pooled regression estimators,

iii) Pooled estimators from intra-class correlation and regression,
iv) Maximum likelihood.

We shall compare the variances of the alternative estimators, together with the sampling correlations between estimates obtained from the same data, using the methods described in section 2.

i) *Intra-class correlation between half-sibs* (H_{ts})

The intra-class correlation between half sibs, H_{ts}, is too well known to require definition here. The approximate sampling variance, modified from Osborne and Patterson [1952] or Robertson [1959], is

$$V(H_{ts}) \doteq \frac{1}{8sd^2n^2}\{(4 - H)^2[4 - 2H - 4K + n(H + 4K) + ndH]^2 + [4 + (d - 1)H]^2$$

$$\times [4 - 2H - 4K + n(H + 4K)]^2/(d - 1) + 4d(n - 1)H^2(2 - H - 2K)^2\} \qquad (12)$$

where the variances deriving from the mean squares for sires, dams and individuals are shown in order. The method can, of course, be used for sex limited traits.

ii) *Regression of offspring on parent performance*

Each of the following regression estimators, not necessarily an exhaustive list, can be shown to be unbiased for H.

a. *Progeny on dam within sires* (H_{bd}). The estimator,

$$H_{bd} = 2\sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})(\bar{Z}_{ij.} - \bar{Z}_{i..})/\sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 \qquad (13)$$

makes no use of differences between sires, and is the typical daughter-dam regression technique used for traits expressed only in females, such as milk yield in cattle where there is often only one daughter for each dam ($n = 1$). From regression theory,

$$V(H_{bd}) \doteq \frac{4 - 2H + nH(1 - H) + 4(n - 1)K}{s(d - 1)n} \qquad (14)$$

and we can show that

$$\text{cov}\,(H_{bd}, H_{ts}) \doteq -\frac{H}{2sd(d - 1)n}[4 + (d - 1)H][4 - 2H + nH(1 - H) + 4(n - 1)K].$$

The regression of H_{ts} on H_{bd} is simply $-H[4 + (d - 1)H]/2d$, but the correlation of the two estimates has a lengthy formula. The correlation is negative if $H > 0$, in contrast to that between the estimates from covariance of full sibs and offspring on mid-parent regression described earlier. Presumably a sample of dams with a genetic variance above expectation induces a regression above average and a sire variance component, estimated from the difference between sire and dam mean squares, below average. Since $V(H_{bd})$, $V(H_{ts})$ and cov (H_{bd}, H_{ts}) are all inversely proportional to s, (under our assumptions) the correlation does not depend on s. Also, if d and s are large, $n = 1$ and $H > 0$, it can be shown that the correlation between H_{bd} and H_{ts} approaches $-H/(2d)^{1/2}$.

b. *Progeny on sire* (H_{bs}). The estimator,

$$H_{bs} = 2 \sum_i (X_i - \bar{X}_.)(\bar{Z}_{i..} - \bar{Z}_{...})/\sum_i (X_i - \bar{X}_.)^2$$

can be used for traits expressed only in males, since it makes no use of information on the dams. We can show that

$$V(H_{bs}) \doteq \frac{4 - 2H + nH + ndH(1 - H) + 4(n - 1)K}{sdn} \tag{15}$$

and

$$\text{cov}\,(H_{bs}\,, H_{ts}) \doteq \frac{H(4 - H)}{2sdn}\,[4 - 2H + nH + ndH(1 - H) + 4(n - 1)K].$$

Thus the regression of H_{ts} on H_{bs} is $H(4 - u)/2$ and, like the correlation, does not decrease to zero as the size of the experiment increases.

c. *Progeny on sire plus dam average* (H_{ba}). The information available on the mean performance of dams mated to each sire is excluded from the regressions H_{bd} and H_{bs}. It can be incorporated by regressing the mean performance of progeny in a sire family on the sire plus average dam performance. Thus

$$H_{ba} = 2 \sum_i (X_i + \bar{Y}_{i.} - \bar{X}_. - \bar{Y}_.)(\bar{Z}_{i..} - \bar{Z}_{...})/\sum_i (X_i + \bar{Y}_{i.} - \bar{X}_. - \bar{Y}_.)^2$$

and

$$V(H_{ba}) \doteq \frac{4 - 2H + n(d + 1)H(1 - H) + 4(n - 1)K}{s(d + 1)n} \tag{16}$$

which is slightly less than $V(H_{bs})$. Also

$$\text{cov}\,(H_{ba}\,, H_{ts}) \doteq \text{cov}\,(H_{bs}\,, H_{ts}) - H^3(4 - H)/2sd.$$

d. *Progeny on mid-parent* (H_{bm}). If the hierarchical structure is disregarded, a straightforward regression of offspring on mid-parent can be computed in which the sire performance is included with each of his mates. It is a simple method of utilizing all the observations on the parents for traits expressed in both sexes, and

$$H_{bm} = 2 \sum_i \sum_j (X_i + Y_{ij} - \bar{X}_. - \bar{Y}_.)(Z_{ij.} - \bar{Z}_{...})/\sum_i \sum_j (X_i + Y_{ij} - \bar{X}_. - \bar{Y}_.)^2.$$

The error structure of this estimator is more complicated since the errors about regression of dam families in the same sire family are correlated, but when s is large the variance reduces to

$$V(H_{bm}) \doteq \frac{4[2 - H + nH(1 - H) + 2(n - 1)K] + (d - 1)nH(1 - H)}{4nsd}. \tag{17}$$

The covariance between H_{bm} and H_{ts} is not required in our subsequent analysis.

e. *Pooled regression estimators* (H_{bp}). It can be shown that H_{bd} (from within sire families) is uncorrelated with both H_{bs} and H_{ba} (from between sire families). For a trait which is expressed in both sexes, it seems reasonable to assume that H_{bd} and H_{ba} contain all the information which can be obtained by regression. From these a pooled estimator, H_{bp}, can be obtained by substituting into (8), (9), and (10), but they simplify such that

$$\alpha = V(H_{ba})/[V(H_{bd}) + V(H_{ba})]$$

and

$$V(H_{bp}) = [1/V(H_{bd}) + 1/V(H_{ba})]^{-1},$$

since H_{bd} and H_{ba} are uncorrelated.

In limiting cases of family size, several of these regression estimators are the same. If $d = 1$ (i.e. one dam per sire), then $H_{bm} \equiv H_{ba} \equiv H_{bf}$ (the latter refers to full sib families, see section 2) and since there is no information on H_{bd}, it follows that $H_{bp} \equiv H_{bm}$ also. Our formulae are not precise if $s = 1$ (only one sire family), but it follows that there is no information on either H_{bs} or H_{ba} and $H_{bd} \equiv H_{bm} \equiv H_{bp}$.

iii) *Pooled estimators from covariance of half sibs and regression*

Estimators can also be obtained by pooling those from the covariance of half sibs and from one or more regression estimators. The appropriate method will depend on whether or not the trait is sex limited. For traits expressed only in males we define H_{ps}, which is a linear function of H_{ts} (from the covariance of half sibs) and H_{bs} (from the regression of progeny on sire). The optimal weighting and $V(H_{ps})$ are based on (8), (9), and (10). For traits expressed only in females, we define H_{pd}, which is a linear function of H_{ts} and H_{bd} (from the regression of progeny on dam), obtained by the same weighting procedure. If a trait is expressed in both sexes, we have suggested that all information from regression is included in H_{bd} and H_{ba}, and these can be combined with H_{ts} to form a pooled estimate H_{pa}, given by

$$H_{pa} = \alpha_1 H_{bd} + \alpha_2 H_{ba} + \alpha_3 H_{ts}$$

with $\sum_i \alpha_i = 1$ and the α_i chosen to minimize $V(H_{pa})$. The solution can be shown to be as follows. Let c_{ij} be the covariance between estimates i and j, and let $A = c_{12} - c_{13} - c_{23} + c_{33}$, $B = c_{11} - 2c_{13} + c_{33}$ and $C = c_{22} - 2c_{23} + c_{33}$. Then

$$\alpha_1 = [(c_{33} - c_{23})A - (c_{33} - c_{13})C]/[A^2 - BC]$$

$$\alpha_2 = (c_{33} - c_{13} - \alpha_1 B)/A, \alpha_3 = 1 - \alpha_1 - \alpha_2 .$$

Of course, only estimates of the c_{ij} are available, so exact weightings are not possible.

The sampling variances of these estimators are compared with those expected from ML methods in section 4.

iv) *Maximum likelihood* (H_m)

Consider the model in which observations are available on both sexes, so that a total of $s + sd + sdn$ measurements are made with the variance-covariance structure given by (11). However, as in the full sib case, it is useful to transform the observations into the following order for each sire family, say sire i:

$$X_i, \bar{Y}_{i\cdot}, \bar{Z}_{i\cdot\cdot}, Y_{i1} - \bar{Y}_{i\cdot}, \bar{Z}_{i1\cdot} - \bar{Z}_{i\cdot\cdot}, \cdots, Y_{i,d-1} - \bar{Y}_{i\cdot}, \bar{Z}_{i,d-1,\cdot} - \bar{Z}_{i\cdot\cdot},$$

$$Z_{i11} - \bar{Z}_{i1\cdot}, \cdots, Z_{i1,n-1} - \bar{Z}_{i1\cdot}, Z_{i21} - \bar{Z}_{i2\cdot}, \cdots, Z_{id,n-1} - \bar{Z}_{id\cdot}.$$

Let this set of observations have variance-covariance matrix $\mathbf{W}_i\sigma^2$, of dimension $1 + d + dn$. Since $\mathbf{W}_i$ is the same for all i, and sire families are uncorrelated, the overall variance-covariance matrix $\mathbf{W}$ of dimension $s + sd + sdn$ is given by

$$\mathbf{W} = \mathbf{I} * \mathbf{W}_i\sigma^2.$$

We have

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{S}_1 & 0 & 0 \\ 0 & \left(\mathbf{I} - \frac{1}{d}\mathbf{J}\right) * \mathbf{S}_2 & 0 \\ 0 & 0 & \mathbf{I} * \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)(1 - H/2 - K) \end{bmatrix}$$

where, in the (2, 2) block of $\mathbf{W}_i$, $(\mathbf{I} - (1/d)\mathbf{J})$ is of dimension $d - 1$, and in the (3, 3) block, $\mathbf{I}$ is of dimension d and $(\mathbf{I} - (1/n)\mathbf{J})$ of dimension $n - 1$. Also

$$\mathbf{S}_1 = \begin{bmatrix} 1 & 0 & \frac{H}{2} \\ 0 & \frac{1}{d} & \frac{H}{2d} \\ \frac{H}{2} & \frac{H}{2d} & \frac{H}{4} + \frac{H}{4d} + \frac{K}{d} + \frac{1 - \frac{1}{2}H - K}{nd} \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} 1 & \frac{1}{2}H \\ \frac{1}{2}H & \frac{1}{4}H + K + (1 - \frac{1}{2}H - K/n) \end{bmatrix}.$$

From the properties of direct products (e.g. Searle [1966]) and utilizing the special form of these "$\mathbf{I} + \mathbf{J}$" matrices (Searle [1970]) we obtain

$$\mathbf{W}^{-1} = \mathbf{I} * \mathbf{W}_i^{-1}$$

$$\mathbf{W}_i^{-1} = \begin{bmatrix} \mathbf{S}_1^{-1} & 0 & 0 \\ 0 & (\mathbf{I} + \mathbf{J}) * \mathbf{S}_2^{-1} & 0 \\ 0 & 0 & \mathbf{I} * (\mathbf{I} + \mathbf{J})/(1 - \frac{1}{2}H - K) \end{bmatrix}$$

and

$$|\mathbf{W}| = |\mathbf{S}_1|^2 \left(\frac{1}{d^2}|\mathbf{S}_2|^{d-1}\right)^s \left[\frac{1}{n}(1 - \tfrac{1}{2}H - K)^{n-1}\right]^{sd}.$$

Hence, the log likelihood can be shown to be

$$\begin{aligned} \text{Log } L = \text{ constant terms } &- \tfrac{1}{2}(s + sd + sdn) \log \sigma^2 - \tfrac{1}{2}s \log |\mathbf{S}_1| - \tfrac{1}{2}s(d - 1) \log |\mathbf{S}_2| \\ &- \tfrac{1}{2}sd(n - 1) \log (1 - \tfrac{1}{2}H - K) - \tfrac{1}{2}\sigma^2 \Bigg[\sum_{i=1}^{s} (\mathbf{x}_i - \mu\mathbf{1})' \mathbf{S}_1^{-1} (\mathbf{x}_i - \mu\mathbf{1}) \\ &+ \sum_{i=1}^{s} \sum_{j=1}^{d} \mathbf{w}_{ij}' \mathbf{S}_2^{-1} \mathbf{w}_{ij} + \sum_{i=1}^{s} \sum_{j=1}^{d} \sum_{k=1}^{n} (Z_{ijk} - \bar{Z}_{ij.})^2 / (1 - \tfrac{1}{2}H - K) \Bigg] \end{aligned}$$

where $\mathbf{x}_i' = (X_i, \bar{Y}_{i.}, \bar{Z}_{i..})$, $\mathbf{1}' = (1, 1, 1)$ and $\mathbf{w}_{ij}' = (\bar{Y}_{ij} - \bar{Y}_{i.}, \bar{Z}_{ij.} - \bar{Z}_{i..})$.

Differentiation of the likelihood and obtaining expectations of the second partial derivatives are straightforward, and the results can be evaluated on a computer. The matrix $\mathbf{P}$, of dimension 4×4, has elements $P_{ij} = -E(\partial^2 \log L / \partial\theta_i \partial\theta_j)$, where we take $\theta_1 = \mu$, $\theta_2 = \sigma^2$, $\theta_3 = H$ and $\theta_4 = K$. The inverse of $\mathbf{P}$ gives the sampling variances and covariances of the ML estimators.

If information is available only on females a total of $s(d + nd)$ observations is available. The sampling variances of the ML estimators are found in the same way, but the first row and column of $\mathbf{S}_1$, together with the relevant terms in the observations which relate to information on sire performance, are deleted. Similarly, if there is no information available on females, there are $s(1 + nd)$ observations, and the second row and column of $\mathbf{S}_1$, the first row and column of $\mathbf{S}_2$ and the appropriate observations are deleted.

When no parental data are available, deletion of the first and second rows and columns of $\mathbf{S}_1$ and the first row and column of $\mathbf{S}_2$ is required. In such a balanced design the estimates of variance components by the analysis of variance are minimum-variance quadratic unbiased (Graybill and Hultquist [1961]) and are equal to the ML estimators after correction for bias with normally distributed observations (Graybill [1954]). Thus the large sample variances of heritability estimates by ML (H_m) and intra-class correlation (H_{t*}) are the same when only progeny data are available.

Pooling of Sheridan et al.'s results

An example of the use of the theory developed in this section can be given by considering the alternative heritability estimates of Sheridan *et al.* [1968]. From the same data they obtained estimates of H_{t*}, H_{bd}, and H_{b*}, for total abdominal and sternopleural bristle number in both male and female *Drosophila melanogaster* with a balanced hierarchical design of $s = 62$, $d = 3$, and $n = 10$. Using the method outlined in section 3(iii), we can obtain a single pooled estimate of heritability for each character in each sex. (In the absence of the original data it has not been possible to pool the male and female estimates, nor has it been possible to obtain an ML estimate). We do this by first guessing a value for the pooled heritability which is then substituted as H into the equations for $V(H_{bd})$, $V(H_{b*})$, $V(H_{t*})$, cov (H_{bd}, H_{t*}), cov (H_{b*}, H_{t*}), and cov (H_{bd}, H_{b*}). The values thus obtained are substituted into the equations for α_1, α_2, and α_3 to provide estimates of these three weights which are then used to obtain a second estimate of pooled heritability as

$$H_p = \alpha_1 H_{bd} + \alpha_2 H_{b*} + \alpha_3 H_{t*} .$$

The cycle is repeated until the estimate of H_p stabilizes.

This final estimate of H_p is then used to obtain final estimates of the expected sampling variances and covariances, and hence the relevant sampling correlation coefficients. The results of these calculations, together with the estimates and standard errors of Sheridan *et al.*, are presented in Table 1.

Each of the pooled estimates is seen to be weighted in favor of the separate estimates with lowest variance, and the standard error of each of the pooled estimates is lower than any of those of the separate estimates, as we would expect. The standard errors expected for each separate estimate are in reasonable agreement with those observed. It can also be seen that the expected sampling correlation between H_{b*} and H_{t*} is never greater than 0.47, and that the correlations between H_{bd} and H_{b*}, and H_{bd} and H_{t*} are expected to be zero and slightly negative respectively. In view of these relatively low correlations, we should not necessarily expect close agreement among the estimates.

4. RELATIVE EFFICIENCY OF ESTIMATORS IN A HIERARCHICAL STRUCTURE

The relative magnitudes of the sampling variances of different heritability estimates from the same set of data depend, of course, on the design parameters, n, d, and s, and also the underlying parameters H and K. Thus we can only compare the estimators for

TABLE 1

RESULTS OF ANALYSIS OF DATA OF SHERIDAN *et al.* ON ABDOMINAL AND STERNOPLEURAL BRISTLE NUMBERS IN *D. Melanogaster*

		Total Abdominal		Sternopleural	
		Males	Females	Males	Females
Heritability ± SE*	H_{bd} (1)	0.28 ± 0.09	0.21 ± 0.08	0.18 ± 0.08	0.26 ± 0.08
	H_{bs} (2)	0.22 ± 0.10	0.40 ± 0.15	0.16 ± 0.09	0.18 ± 0.13
	H_{ts} (3)	0.29 ± 0.13	0.67 ± 0.18	0.17 ± 0.08	0.29 ± 0.10
Pooled H		0.26 ± 0.062	0.35 ± 0.065	0.17 ± 0.046	0.25 ± 0.050
Expected SE	(1)	0.093	0.094	0.072	0.074
	(2)	0.097	0.102	0.078	0.085
	(3)	0.123	0.139	0.076	0.092
Expected Correlations	(1,2)	0.00	0.00	0.00	0.00
	(1,3)	-0.15	-0.19	-0.12	-0.15
	(2,3)	+0.39	+0.47	+0.34	+0.43

* Calculated by Sheridan *et al.*

a few examples. All but one of the designs have been chosen such that for an intermediate H value (0.2), they are the optimum for ML estimation, given a fixed total number scored, T. The single exception is the design used for the comparison of estimators in Figure 3a. This design, which is far from optimum, has been chosen to illustrate that the conclusions drawn from comparisons are quite robust over different designs.

The results are given in Figures 3, 4, and 5 for traits in which both sexes are scored, only females are scored and only males are scored, respectively. In each case variances are expressed on a single observation basis, i.e. they are the inverses of the Fisherian information per observation. A large number of sires is assumed to be used, so that the variance of each estimator is inversely proportional to the number of sires. This assumption is less satisfactory for estimators such as the regression of progeny on sire (H_{bs}) or the half sib intra-class correlation (H_{ts}), for with only one sire available H_{bs} and H_{ts} cannot be estimated. Then the only unbiased information on heritability comes from the regression of progeny on dam (H_{bd}), so the ML estimator (H_m) must then have the same efficiency.

In Figure 3 and in other examples we have investigated in which the estimators can be compared, it is seen that H_{bd} has a considerably lower variance than the other single parent regression estimator H_{bs} . Also H_{ba} , the regression on sire and dam average, has a variance intermediate between the single parent regression estimators over most heritability values. The regression on mid parent, H_{bm} , is more efficient than H_{bd} . The only intra-class correlation estimator which is unbiased, H_{ts} , may be more efficient than any regression estimator at low heritabilities, but becomes very much worse at high heritabilities. This was shown for some of the estimators by Robertson [1959]. The variance of the ML estimator, H_m , is much smaller than that of the best commonly used estimator, but the pooled estimators, H_{bp} , based only on regression estimators and H_{pa} , based on all estimators, are not much less efficient than H_m . At low heritabilities H_{pa} and H_m have almost the same sampling variance.

A few assumptions need to be emphasized, however: the exact weightings for the pooled estimates could not be achieved exactly, so the designs have been chosen to be near optimal for ML estimation without regard to their efficiency for other estimators. The variances are expressed in terms of all the observations in the experiment, $s + sd + sdn$ assuming these have all been obtained. However, H_{ts} is based on sdn observations, H_{bs} on $s + sdn$ and H_{bd} on $sd + sdn$. Therefore the variances of each of these estimators per observation required for them are also shown in Figure 2 (denoted H_{ts}', H_{bs}', and H_{bd}', respectively). The variance of the half-sib covariance estimator is the one most affected by this modification, especially when family sizes are small, but it remains less efficient than ML based on parental and progeny information except with very low heritabilities and some family size combinations.

For sex limited traits scored only in females (Figure 4), the pooled estimator H_{pd} is considerably more efficient than the simple regression estimator H_{bd} and is as efficient as ML at low heritabilities. At higher heritabilities H_{pd} is little better than H_{bd} and somewhat

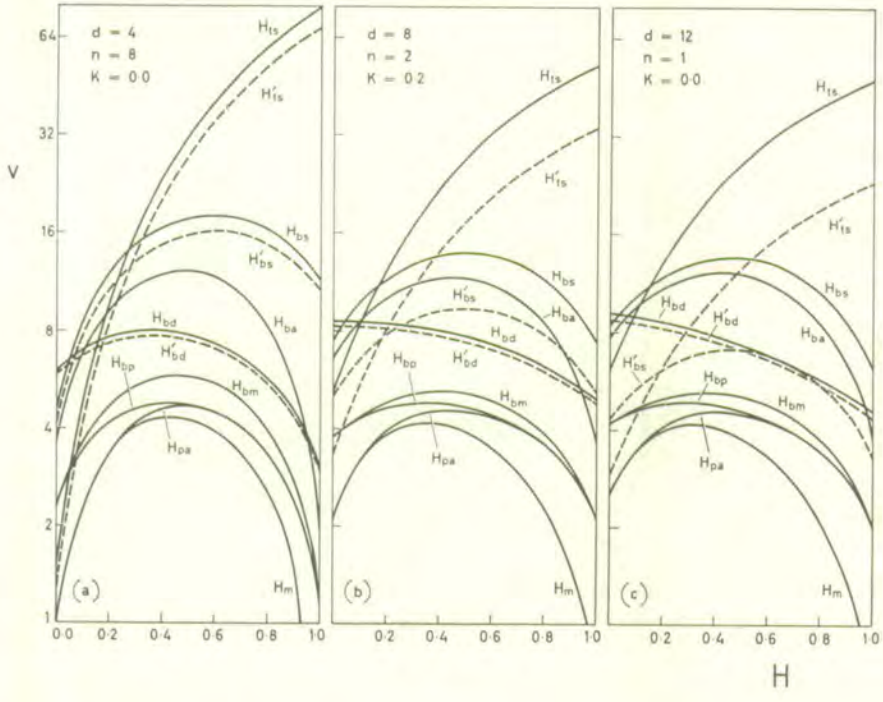

FIGURE 3

SAMPLING VARIANCES PER OBSERVATION (v) OF ALTERNATIVE HERITABILITY ESTIMATORS FOR TRAITS MEASURED IN BOTH SEXES USING FULL AND HALF SIB FAMILIES WITH DATA COLLECTED ON PARENTS AND PROGENY (SOLID CURVES): REGRESSION ON SIRE PERFORMANCE (H_{bs}), ON SIRE AND MEAN DAM PERFORMANCE (H_{ba}), ON DAM PERFORMANCE (H_{bd}), ON MID-PARENT (H_{bm}), A POOLED REGRESSION ESTIMATOR (H_{bp}), FROM HALF-SIB COVARIANCE (H_{ts}), A POOLED REGRESSION AND SIB COVARIANCE ESTIMATOR (H_{pa}) AND MAXIMUM LIKELIHOOD (H_m). THE CORRESPONDING SAMPLING VARIANCE PER OBSERVATION REQUIRED (BROKEN LINES) FOR THOSE ESTIMATORS NOT USING ALL DATA ARE ALSO GIVEN (H_{bs}', H_{bd}', H_{ts}').

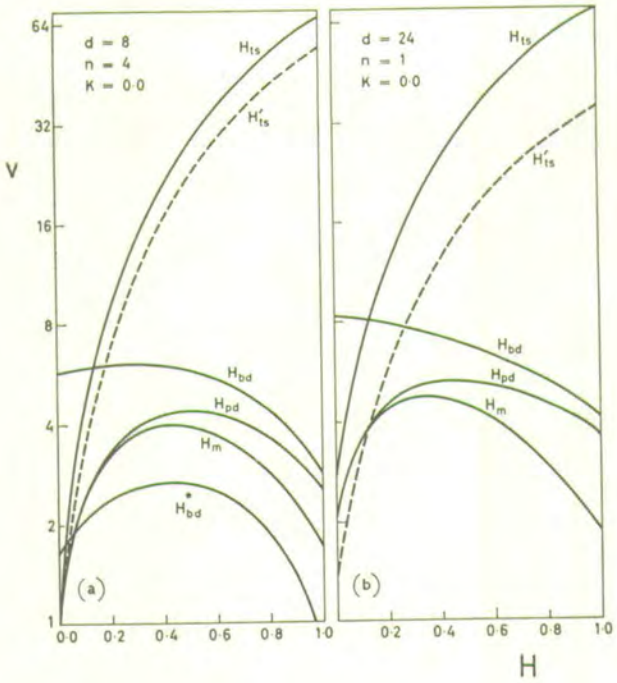

FIGURE 4

As Figure 3, but for traits recorded only on females, together with H_{bd}^*, estimated from regression on selected parents

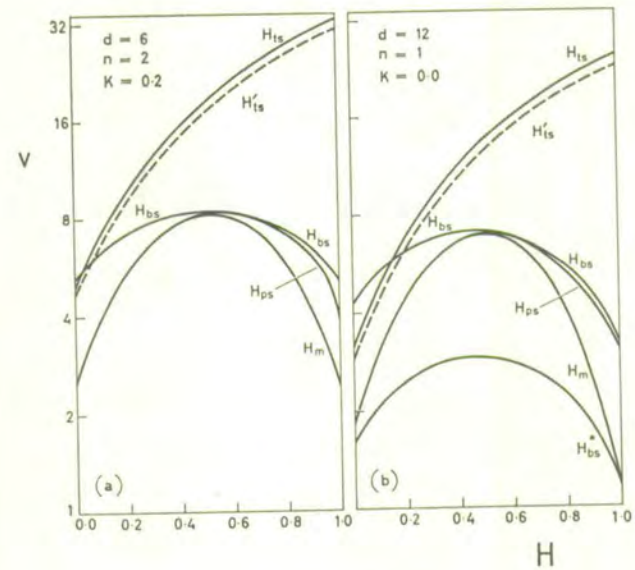

FIGURE 5

As Figure 3, but for traits recorded only on males, together with H_{bs}^*, estimated from regression on selected parents

poorer than H_m . When the trait is scored only in males (Figure 5) similar conclusions hold for the regression estimator H_{bs} rather than H_{bd} , and the pooled estimator H_{ps} rather than H_{pd} .

In Figures 3, 4, and 5 examples are also given for designs in which only half-sib data is available (i.e. $n = 1$). In these it is assumed that $K = 0$, since there are no full sib families from which it can be estimated. The general patterns are seen to be very similar to those of the relevant full hierarchical structure shown in the same figure.

As well as providing comparisons of efficiency of various heritability estimators, Figures 3, 4, and 5 also provide information of potential use in the planning of experiments to estimate heritability. Given an optimum sire family design, Figures 3, 4, and 5 can then be used to provide a direct indication of the total number of observations required to achieve an estimate of heritability with a particular variance. Suppose, for example, that we wished to obtain an estimate of H_{bm} with a standard error of 0.1 for a character in which we expect both the heritability and K to be around 0.2. Using Table 2, the optimum values of d and n are 8 and 2 respectively, and from Figure 3b, we see that $v \doteq 5$ for H_{bm} at $H = 0.2$ with this design. Since $v = T \cdot V(H_{bm})$ and $T = s[1 + d(n + 1)] = 25s$ in this case, we have $V(H_{bm}) = 5/25s$. But we want $V(H_{bm}) = 0.01$ which therefore requires $s = 5/25 \times 1/0.01 = 20$ sire families or a total of 500 observations over the two generations. More generally, a similar type of conclusion can be obtained by the use of the relevant equation in section 3, for any commonly used heritability estimator and for any particular combination of H, K, d, and n. Again it should be noted that such a conclusion will often be quite robust for a range of values of the parameters H and K. In Figure 3b for example, it can be seen that our conclusion for $H = 0.2$ would equally apply to all values of H between 0.2 and 0.6.

Some indication of the probable value of K may be available from previous analyses, as is often the case with heritability. In terms of the model of section 3, we have $K = (\frac{1}{4}V_D + V_{Ec})/V_P$, using the notation of Falconer [1960]. An indication of its probable value can therefore be obtained as $\hat{K} = (H_{id} - H_{is})/4$, where H_{id} is the half-sib heritability estimate based on the dam component of variance. Such an estimate must of course be interpreted with considerable caution, because of sampling errors involved in estimating H_{id} and H_{is} .

The optimum values of d and n for use in calculations such as those just outlined have been determined by Robertson [1959] for intra-class correlation estimates and by Latter and Robertson [1960] for regression estimates. Now that we have an expression for $V(H_m)$, we can examine the relative efficiencies of different experimental designs for ML estimation of heritability, and compare these optimum values of d and n with those relevant to the regression and intra-class correlation estimates.

5. OPTIMUM DESIGNS FOR HERITABILITY ESTIMATION

We now find optimum designs for ML estimation using both parent and progeny data, making the same assumptions as Robertson [1959] and Latter and Robertson [1960] of random mating among unselected parents. It has not proved possible to find the optimum designs for ML analytically so our results have been obtained by trial and error numerical evaluation of $V(H_m)$ on a computer. In all cases we define the optimum design as that giving the most information, i.e. $V(H_m)^{-1}$, per observation on either parent or progeny. Since the large sample variance of H_m that we have to use is inversely proportional to s (the number of sires) the optimum design depends only on d and n.

TABLE 2

OPTIMUM FAMILY STRUCTURE (d, n) FOR MAXIMUM LIKELIHOOD ESTIMATION OF HERITABILITY IN A HIERARCHICAL DESIGN WITH PARENTS AND PROGENY SCORED

K	Sexes Scored	H				
		0.05	0.10	0.20	0.40	0.60
0.00	♂ & ♀	11,8	8,5	5,4	3,3	3,2
	♂	36,2	16,2	6,2	2,2	2,2
	♀	11,9	10,5	8,4	7,3	9,2
0.05	♂ & ♀	22,4	13,3	6,3	3,2	3,2
	♂	38,2	16,2	6,2	2,2	2,2
	♀	25,4	23,2	13,2	9,2	9,2
0.20	♂ & ♀	44,2	20,2	8,2	4,2	3,2
	♂	43,2	18,2	6,2	2,2	2,2
	♀	50,2	27,2	15,2	11,2	10,2

For the hierarchical structure analyzed in section 3, the optimum designs for ML estimation are given in Table 2 for a range of values of H and K, and for characters measured either in both sexes or in males or females alone. The optimum values of d increase if there is a decrease in H or an increase in K. A similar trend is observed in n at low K, but as the covariance between full sibs becomes increasingly inflated by maternal environment or nonadditive genetic effects, the optimum value of n soon reduces to 2, which is the lowest value of n for which K can be estimated. For characters scored only in males, the optimum design does not depend greatly on K, and at higher H values is close to the optimum design for traits measurable in both sexes. Only at high heritabilities does the optimum design for traits measured just in females differ greatly from that appropriate for both sexes. Thus it should be possible to select a design which provides a high degree of efficiency for the simultaneous estimation of heritability of several sex-limited and nonsex-limited traits. Table 2 shows, however, that it is more difficult to find a suitable compromise for traits of widely differing heritability or maternal environment correlation. It can be seen in Table 2 that, for constant H, the optimum value of nd does not depend greatly on K. With both sexes scored, these optima are roughly 88, 40, 18, 7, and 6 for $H = 0.05$, 0.1, 0.2, 0.4, and 0.6 respectively. As a good approximation, the value of nd at the optimum is $4/H$, giving $nd = 80, 40, 20, 10$, and 7 respectively. If only males are scored, the optimum for nd is $3/H$ approximately, and if only females are scored it becomes $5/H$ approximately.

These results do not differ greatly from those derived by Robertson [1959] for heritability estimation from the covariance of half sibs. He found that a dam family size (n) of one with $d = 4/H$, approximately, to be the optimum. If both sire and dam intra-class correlations are to be estimated Robertson showed that the optimum value of n was $2/H$, with $d = 3$ or 4. These values of n are slightly larger and d slightly smaller than those given in Table 2 for ML estimation using both parental and progeny data. As we have noted previously, the half sib intra-class correlation estimator and the ML estimator are essentially the same when only progeny data are available, and so therefore are their respective optimum designs.

The optimum designs have also been found by computation for cases in which both parents and progeny are measured, but where only half sib families (i.e. $n = 1$) are available in the progeny generation. A value of $K = 0$ has been assumed since it can not be estimated.

TABLE 3

OPTIMUM HALF-SIB FAMILY SIZE (d) FOR MAXIMUM LIKELIHOOD ESTIMATION OF HERITABILITY WHERE OBSERVATIONS ARE AVAILABLE ON PARENTS AND HALF-SIB PROGENY ONLY

Sexes Scored	H				
	0.05	0.10	0.20	0.40	0.60
♂ & ♀	71	31	12	5	4
♂	70	30	10	4	4
♀	82	43	24	15	14

The results are shown in Table 3, and it is seen that the optimum value of d (and hence nd) is generally somewhat smaller than the optimum value of nd when both full and half sibs are available (Table 2). If only full sib families are available the optimum design if K is to be estimated is close to that given by Latter and Robertson [1959], presumably since all information on H comes from regression of offspring on parent.

Many of the optimum designs shown in Tables 2 or 3 may be impracticable, especially those requiring large values of d. However, apparently large departures from the optimum design often involve only a small reduction in the amount of information per observation. Some examples to illustrate this are given in Figure 6; similar results have been found for

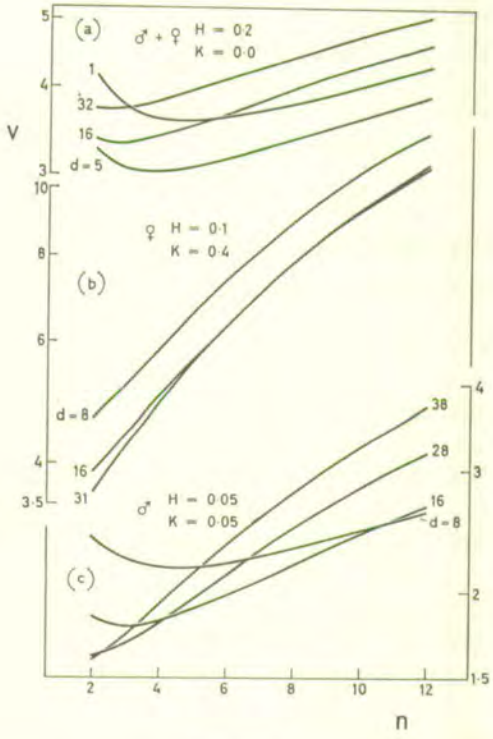

FIGURE 6

SAMPLING VARIANCE PER OBSERVATION (v) OF ML ESTIMATORS OF HERITABILITY FOR DIFFERENT FAMILY SIZES, WITH RECORDS ON BOTH SEXES (a), ONLY FEMALES (b), OR ONLY MALES (c).

other combinations. We see that for a trait scored only in females with a low H and high K, a reduction in d from the optimum of 31 down to 16 increases the variance per observation by only 6% if n remains at 2.

Although Tables 2 and 3 give the optimum designs when there is prior knowledge of H and K, there is also need to specify designs likely to be efficient over a wide range of parameter values when this prior knowledge is absent. We find that a satisfactory design has a dam family size (n) of 2, and 6 dams per sire (d) for characters scored in both sexes or in males alone and 12 dams per sire for characters scored only in females. If only parental and half sib information is available ($n = 1$), then the optimum number of dams per sire is around 12 and 24 respectively. When only parental and full sib data are available ($d = 1$), a full sib family size of 3 is efficient over a wide range of parameters.

6. DISCUSSION

Let us first review our more important assumptions and consider their implications. The omission of a term for dominance or common environment (K) in the full sib model was made primarily to enable simpler demonstration of the principles; it can not be defended too strongly in practice. We also ignored any environmental covariance of dam and offspring in the hierarchical case. Such covariances certainly exist, for example in litter size in mice (Falconer [1955]). It would not be difficult to include such a term in the model; then all the unbiased information on heritability would come from the regression of progeny on sire (only for traits expressed in males) and the covariance of half sibs, whose properties have been analyzed in section 3. The assumptions of equality of means and variances in the two generations and sexes are likely to have biased the sampling variances downwards, but few degrees of freedom would be lost in their estimation. Experiments from which heritability estimates are obtained are rarely balanced, except perhaps in *Drosophila*. Removal of this assumption should introduce no conceptual difficulties in ML estimation, but would make the form of the variance-covariance structure of the alternative regression and sib covariance estimators rather involved. The mechanics of the ML estimation procedure have not been considered, but a specific program for this sort of data has been written (Felsenstein, personal communication) and there are many general programs for finding maxima.

Throughout we have assumed that there is no selection or assortative mating of the parents, yet both can give much reduced sampling variances of regression estimators in a properly designed experiment (Hill [1970]). Two examples are given for sex-limited traits in Figures 4(a) and 5(b), with the optimum designs appropriate for selection of parents with $H = 0.2$, the same value used to choose the design for ML estimation. In Figure 4(a) we have used $n = 14$ and a proportion of 5.5% of potential female parents selected (from Hill [1970]). This estimator of regression of progeny on selected parents, $H_{bd}{}^*$, has a variance approximately half of the ML estimator, H_m, per individual scored, except at very low heritabilities. Similarly, in Figure 5(b), selection of males gives an estimator, $H_{bs}{}^*$, with substantially lower sampling variance than H_m, particularly at intermediate heritabilities. Thus where selection can be practiced, we advocate that it be done. Even then there will be some information available from the variance between families. ML methods which could deal with such data have been developed by Thompson [1973].

There are several situations where selection or assortative mating of the parents may not be desirable, however. One such case is a control population being maintained for several generations alongside selected populations to establish whether trends are genetic

or environmental. Usually no selection is practiced in these, but if selection or assortative mating were practiced in a control, it would be to reduce rather than inflate the variance between parents (Hill [1972]) and would reduce the efficiency of heritability estimators. The other main case where neither selection nor assortative mating is desirable is where heritabilities and genetic correlations are to be estimated simultaneously on several traits.

We make two essential recommendations. First, people obtaining estimates of heritability by several methods from essentially the same set of data should take note of the correlation structure among their estimates before concluding that agreement between them is good or bad. Second, all available data should be used to obtain a single estimate; we have considered just pairs of generations, but in a control population several generations might be combined.

ACKNOWLEDGMENT

We are very grateful to Dr. J. Felsenstein who originally suggested the problem to us, to Mr. R. Thompson for noting an error in an earlier version, and to both of them and other colleagues for many helpful comments.

ESTIMATION DE L'HERITABILITE PAR REGRESSION DES DESCENDANTS SUR LEURS PARENTS AINSI QUE PAR CORRELATION INTRA CLASSE SUR DES FRERES DANS UNE EXPERIENCE

RESUME

On discute l'analyse et le plan d'expérience afin d'estimer l'héritabilité quand on se sert de données sur les parents et sur les descendants. On montre qu'il y a une corrélation positive importante entre la régression du descendant sur le parent moyen et la covariance entre pleins frères estimées sur les mêmes données; que dans une structure hiérarchique la covariance entre demi-frères a une corrélation négative avec la régression du descendant sur la mère et une corrélation positive avec la régression du descendant sur le père.

Les efficacités des estimateurs de l'héritabilité, par régression, par covariance entre frères, par combinaison des deux anisi que ceux du maximum de vraisemblance sont comparés. L'estimateur du maximum de vraisemblance ne réduit pas la variance de beaucoup, relativement aux estimateurs combinés mais, chacun d'eux est souvent bien meilleur que les estimateurs obtenus par régression ou par la covariance entre frères.

On décrit les plans d'expérience optimaux pour l'estimateur du maximum de vraisemblance. On trouve qu'ils ne diffèrent pas beaucoup de ceux qui sont appropriés ou pour l'estimation de la régression de descendants sur les parents ou pour l'estimation de la covariance entre demi-frères; et que les plans optimaux sont franchement robustes à tout changement d'hypothèses sur les paramètres.

REFERENCES

Clayton, G. A., Morris, J. A. and Robertson, A. [1957]. An experimental check on quantitative genetical theory. I. Short term responses to selection. *J. Genet. 55*, 131–51.

Falconer, D. S. [1955]. Patterns of response in selection experiments with mice. *Cold Spr. Harb. Symp. Quant. Biol. 20*, 178–96.

Falconer, D. S. [1960]. *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh.

Fisher, R. A. [1925]. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Graybill, F. A. [1954]. On quadratic estimates of variance components. *Ann. Math. Statist. 25*, 367–72.

Graybill, F. A. and Hultquist, R. [1961]. Theorems concerning Eisenhart's Model II. *Ann. Math. Statist. 32*, 261–9.

Hill, W. G. [1970]. Design of experiments to estimate heritability by the regression of offspring on selected parents. *Biometrics 26*, 566–71.

Hill, W. G. [1972]. Estimation of genetic change. I. General theory and design of control populations. *Anim. Breed. Abstr. 40*, 1–15.

Kendall, M. G. and Stuart, A. [1973]. *The Advanced Theory of Statistics, Vol. 2. Inference and Relationship.* Griffin, London.

Latter, B. D. H. and Robertson, A. [1960]. Experimental design in the estimation of heritability by regression methods. *Biometrics 16*, 348–53.

Osborne, R. and Patterson, W. S. B. [1952]. On the sampling variance of heritability estimates derived from variance analyses. *Proc. Roy. Soc. Edinb., B, 64*, 456–61.

Robertson, A. [1959]. Experimental design in the evaluation of genetic parameters. *Biometrics 15*, 219–26.

Searle, S. R. [1966]. *Matrix Algebra for the Biological Sciences.* Wiley, New York.

Searle, S. R. [1970]. Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics 26*, 505–24.

Sheridan, A. K., Frankham, R., Jones, L. P., Rathie, K. A. and Barker, J. S. F. [1968]. Partitioning of variance and estimation of genetic parameters for various bristle number characters of *Drosophila melanogaster*. *Theor. Appl. Genet. 38*, 179–87.

Tallis, G. M. [1959]. Sampling errors of genetic correlation coefficients calculated from the analyses of variance and covariance. *Aust. J. Stat. 1*, 35–43.

Thompson, R. [1973]. The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics 29*, 527–50.

Received January 1973, *Revised October* 1973

Key Words: Heritability estimation; Maximum likelihood estimation; Sampling covariances of regression and intra-class correlation estimators; Design of genetic experiments.

Reprinted from
BIOMETRICS Copyright © 1974
THE BIOMETRIC SOCIETY, Vol. 30, No. 3, September 1974

10

Design and efficiency of selection experiments for estimating genetic parameters

by

William G. Hill

Reprinted from
Biometrics *27*, 293–311
June 1971

DESIGN AND EFFICIENCY OF SELECTION EXPERIMENTS FOR ESTIMATING GENETIC PARAMETERS

William G. Hill

Institute of Animal Genetics, Edinburgh EH9 3JN, Scotland

SUMMARY

(i) Formulae are derived for the sampling variance of selection response and for estimates of realised heritability and realised genetic correlation. (ii) If a control population is maintained, or divergent selection practised, the greater part of the sampling variance comes from genetic drift and depends primarily on the total number of individuals recorded in the whole experiment, rather than on its duration. (iii) The optimal selection intensity for estimating realised heritabilities is investigated—proportions selected of about 15 per cent should be satisfactory. Similar designs will also be efficient for estimation of realised genetic correlations. (iv) Several methods of estimating heritability are compared, of these the realised heritability has least variance. (v) Some selection indices for improving a single trait are evaluated. Mass selection is likely to be best for comparing response from alternative selection programmes or populations.

INTRODUCTION

Selection experiments can be used to estimate heritabilities or other genetic parameters in a population and to compare responses under alternative selection schemes. It is therefore essential to have some information on the precision of the estimates obtained. Much of the variability in response comes from genetic sampling or drift and account must be taken of this in any predictions. Formulae for the variance in gain from a single generation of selection have been given by Prout [1962]. His results were extended to experiments of several generations duration by Soller and Genizi [1967], but they appear to have done so incorrectly. A different result is derived in this paper and is used to compute the variance of an estimate of realised heritability (cf. Falconer [1960]). The formulation is extended to include the variance of response in correlated traits, and an approximate sampling variance is derived for the realised genetic correlation (Falconer [1960]). The implication of these results on the design of selection experiments is then discussed, and comparisons made of the relative efficiency of selection experiments and standard offspring-parent regression or half-sib correlation techniques for estimating genetic parameters.

Several simplifying assumptions are made in the model. Of these the most restrictive is that the genetic variances and covariances do not change within the population during selection. Therefore we have to assume that individual genes each have a small effect on the quantitative trait, that

there are no confounding linkage effects, and that the total inbreeding does not become high. Similar assumptions were made by Soller and Genizi [1967]. Further, we generally assume that during the selection experiment a control population is maintained or response compared in two concurrent lines so that all trends or random deviations caused by environmental changes common to all individuals are removed.

Realised heritabilities will be predicted only from the ratio of total gain to total selection differential up to the last generation of selection. Some information is contained in the means of earlier generations, but will be ignored here. Falconer [1960] and Richardson *et al.*, [1968] utilise some of this information by fitting a linear regression to cumulative response and cumulative selection differential each generation. But with genetic sampling (drift) the variance of the population mean increases each generation, and these means become correlated. In standard regression analysis the observations are assumed to have equal variance and be uncorrelated, so that the estimates of variance of realised heritability obtained by Falconer or Richardson *et al.* using standard regression techniques are biassed downwards. In other words, the observed variance among heritability estimates from a replicated experiment would exceed the variance predicted from a single replicate. In fact, less information is wasted by using only the mean in the last generation than might be expected, and when all the variance is contributed by drift the sampling variances obtained from this technique can be shown to be close to those obtained by estimators using all the information. However, the general problem of estimation of realised heritabilities will form the subject of another paper, which will be concerned primarily with the analysis of experimental data. More emphasis will be given here to problems of design and efficiency under simplified assumptions.

VARIANCE OF RESPONSE

Let us consider the following idealised selection experiment for some quantitative trait with additive gene action in a monecious species. At generation 1 measurements X_{11} , $\cdots$, X_{1M} , with mean $\bar{X}_1$, are taken on a group of M individuals in a closed random mating population with mean $\bar{Z}_0$. From these a group of size N ($\leq M$) are selected to be parents of the next generation, and these would typically be the N individuals with the highest scores for the trait under selection. Let the measurements on the phenotypes of these individuals be Y_{11} , $\cdots$, Y_{1N} , with mean $\bar{Y}_1$, and let their genotypic values be Z_{11} , $\cdots$, Z_{1N} , respectively, with mean $\bar{Z}_1$. The observed selection differential at generation 1 is therefore $\bar{Y}_1 - \bar{X}_1$. Under an additive model, the expected performance of the progeny of the selected individuals, if these are mated at random, is $\bar{Z}_1$. Now a new group of M individuals from these matings are reared and recorded, and these have mean performance $\bar{X}_2$. Selection is again practised as in the first generation, and the experiment continued for a total of t selections, so that the final observation is $\bar{X}_{t+1}$. We now wish to find the variance-covariance structure

of the $\bar{X}_i$, conditional on the observed selection differentials, $\bar{Y}_i - \bar{X}_i$.

Phenotypic and genotypic values are assumed to be bivariate normally distributed, with variances σ^2 and $h^2\sigma^2$, respectively, and correlation h. Thus h^2 is the heritability, or regression of genotype on phenotype.

(i) *Genetic drift in a single generation*

The main problem is to compute $V(\bar{Z}_i - \bar{Z}_{i-1} \mid \bar{Y}_i - \bar{X}_i)$, which is the variance of genetic change, or drift variance, in a single generation. By virtue of the underlying genetic sampling process, this occurs independently of genetic drift from previous or subsequent generations. To simplify the notation, let $\mu = \bar{Z}_{i-1}, \bar{Z} = \bar{Z}_i, Z_j = Z_{ij}, \bar{Y} = \bar{Y}_i, Y_j = Y_{ij}$, and $\bar{X} = \bar{X}_i$. Using standard regression theory, and the fact that μ is the expected value of progeny of the subsequent generation, we have

$$Z_j = \mu + h^2(Y_j - \mu) + e_j,$$

where the e_j are independently distributed with $E(e_j) = 0$, $V(e_j) = h^2(1 - h^2)\sigma^2$. Thus

$$\bar{Z} = \mu + h^2(\bar{Y} - \mu) + \bar{e}, \tag{1}$$

where $V(\bar{e}) = h^2(1 - h^2)\sigma^2/N$ is the variance about the predicted mean obtained by Prout [1962]. However, in the selection experiment μ is not known, so the selection differential is estimated by $\bar{Y} - \bar{X}$. Rewriting (1) as

$$\bar{Z} - \mu = h^2(\bar{Y} - \bar{X}) + h^2(\bar{X} - \mu) + \bar{e}$$

we have

$$E_{\bar{e},\bar{X}}(\bar{Z} - \mu \mid \bar{Y} - \bar{X}) = h^2(\bar{Y} - \bar{X}) + h^2E_{\bar{X}}(\bar{X} - \mu \mid \bar{Y} - \bar{X}) + E_{\bar{e}}(\bar{e} \mid \bar{Y} - \bar{X})$$

where the subscript denotes the variable over which the expectation is to be taken. Now, since we assume the M individuals are sampled at random, $\bar{X}$ is a complete sufficient statistic for μ, and as $\bar{Y} - \bar{X}$ does not depend on μ, then $\bar{X} - \mu$ and $\bar{Y} - \bar{X}$ are independently distributed (Basu [1955]). Similarly, the error about regression, $\bar{e}$, is independent of $\bar{Y}$ or $\bar{X}$, so that

$$E_{\bar{e},\bar{X}}(\bar{Z} - \mu \mid \bar{Y} - \bar{X}) = h^2(\bar{Y} - \bar{X}).$$

Using the same independence relationships we have

$$\begin{aligned} V_{\bar{e},\bar{X}}(Z - \mu \mid \bar{Y} - \bar{X}) &= h^4V_{\bar{X}}(\bar{X} - \mu) + V_{\bar{e}}(\bar{e}) \\ &= h^4\sigma^2/M + h^2(1 - h^2)\sigma^2/N \\ &= h^2\sigma^2[1 - (1 - p)h^2]/N, \end{aligned} \tag{2}$$

where $p = N/M$ is the proportion selected.

Two special cases of (2) are of interest. When no selection is practised, $M = N$, and the variance increases to $h^2\sigma^2/N$, the usual formula for drift variance in an unselected population. At the other extreme, as M becomes infinite with N constant, the variance approaches $h^2(1 - h^2)\sigma^2/N$, the formula appropriate for selection from a population of known mean.

Soller and Genizi [1967] give the value for the variance of response in a single generation as $h^2(1 - h^2)\sigma^2(1/M + 1/N)$. They do not include a term for error arising from the use of an estimate of the selection differential, but assume an error about the regression prediction outside the selection process, i.e. $h^2(1 - h^2)\sigma^2/M$, and both assumptions seem unjustified. Their results for response from several generations of selection are consequently incorrect also.

(ii) *Several generations*

We now extend the preceding basic result to include several generations of selection, and find the variance among the observables, $\bar{X}_t$. To make this possible in any general way, we need to assume that the distribution of genotypic and phenotypic values remains bivariate normal with changing means, but with the same variance-covariance structure in each successive generation as in the first. Changes in these parameters can occur either as the gene frequencies alter as a result of genetic drift or selection, or directly from the truncation selection affecting the distribution of genotypic values among selected individuals. Although these factors are not independent, we consider their importance separately, taking the effect of selection on gene frequencies first.

In a simple additive model, determined by n loci in which the homozygotes differ by a_j , $j = 1, \cdots, n$, in genotypic value, and have frequency q_j , the genetic variance is $\sum \frac{1}{2}q_j\ (1 - q_j)a_j^2$ (Falconer [1960]). The selective value of a gene of effect a_j is ia_j/σ, where i is the selection differential in standard deviations (Falconer [1960]). Thus the genetic variance, computed from the gene frequencies, becomes $\sum \frac{1}{2}q_j\ (1 - q_j)a_j^2[1 + ia_jq_j(1 - q_j)(1 - 2q_j)/\sigma]$ in the next generation. So long as gene effects are small, such that $a_j \ll \sigma$, and especially when, on average, gene frequencies are near 0.5, changes in the variance will be small.

The effect of inbreeding, taken separately from selection, is to reduce the genetic variance within the population in an additive model. For an idealised population of size N, the variance at generation t is given by $(1 - 1/2N)^t h^2\sigma^2$ (Falconer [1960]), which for large N and small t, approximates $(1 - t/2N)h^2\sigma^2$. So long as $t/2N$ is small, i.e., the experiment continues only to a low inbreeding coefficient, the change in the variance will be unimportant.

The effect of selection on the distribution of genotypes of selected individuals has been considered by Pearson [1903], Cochran [1951], and Finney [1956; 1961]. With truncation selection, the density function of genotypes of selected individuals is given by

$$f(u) = \frac{1}{p}\phi(u)\Phi[(uh - x)/(1 - h^2)^{\frac{1}{2}}] \qquad -\infty < u < \infty,$$

where, for simplicity, the genetic variance is taken as 1, and x is the abscissa corresponding to selection of a proportion p from a standardised normal with density function ϕ and distribution function Φ. Cochran [1951] notes

that this function is positively skewed to a marked degree if h is high and p small; otherwise the skewness is only moderate and the general appearance is similar to that of a normal curve. For $h^2 \leq 0.6$ and $p \geq 0.05$, the bounds for most cases of practical importance, the degree of skewness is small. Further, the density function of the genotypic values of progeny of selected individuals, in which there is segregation within families that can be assumed to give a normal progeny distribution in the infinitesimal model, will be even closer to normal in form. The phenotypic distribution of progeny will depart less from normality than the genotypic since the environmental component is normal. Thus the assumption of bivariate normality in successive generations seems reasonable.

For the response in several generations we require

$$V(\bar{X}_{t+1} - \bar{X}_1 \mid \bar{Y}_1 - \bar{X}_1, \cdots, \bar{Y}_t - \bar{X}_t) = V(\bar{Z}_1 - \bar{X}_1 \mid \bar{Y}_1 - \bar{X}_1) + \sum_{i=2}^{t} V(\bar{Z}_i - \bar{Z}_{i-1} \mid \bar{Y}_i - \bar{X}_i) + V(\bar{X}_{t+1} - \bar{Z}_t). \quad (3)$$

Now

$$\bar{Z}_1 - \bar{X}_1 = h^2(\bar{Y}_1 - \bar{X}_1) - (1 - h^2)(\bar{X}_1 - \bar{Z}_0) + \bar{e}_1$$

and using the same arguments as before

$$V(\bar{Z}_1 - \bar{X}_1 \mid \bar{Y}_1 - \bar{X}_1) = (1 - h^2)^2\sigma^2/M + h^2(1 - h^2)\sigma^2/N.$$

The variance, $V(\bar{X}_{t+1} - \bar{Z}_t)$, depends on the family structure. Within full sib families the variance is $(1 - \frac{1}{2}h^2)\sigma^2$, and between full sib families the variance comprises two parts: the first is $\frac{1}{2}h^2(1 - h^2)\sigma^2$, which is the genetic variance among the means of pairs of individuals with the same phenotypic value. The second depends on the selection criterion; we shall assume that directional selection is practised. The variance of phenotypes among the selected individuals is $[1 - i(i - x)]\sigma^2$ where i and x are defined above (Pearson [1903]). Hence the genetic variance among pairs of individuals from this source is $\frac{1}{2}h^4[1 - i(i - x)]\sigma^2$, so the total variance between families is $\frac{1}{2}h^2(1 - h^2)\sigma^2 + \frac{1}{2}h^4[1 - i(i - x)]\sigma^2 = \frac{1}{2}h^2[1 - h^2i(i - x)]\sigma^2$. For $p = 0.3$, 0.1, and 0.02, $i(i - x) = 0.74$, 0.83, and 0.89, respectively, i.e., it does not depart far from unity; so as a first approximation we ignore the term contributed from variance among selected individuals, and let the between family variance be $\frac{1}{2}h^2(1 - h^2)\sigma^2$. If all families are of equal size, there is no between family contribution to the variance of progeny means about parental means and

$$V(\bar{X}_{t+1} - \bar{Z}_t) = (1 - \tfrac{1}{2}h^2)\sigma^2/M.$$

If family sizes are Poisson-distributed there is an equal contribution from the between and within family components, and

$$V(\bar{X}_{t+1} - \bar{Z}_t) = (1 - \tfrac{1}{2}h^4)\sigma^2/M$$

with directional selection of at least moderate intensity.

In our earlier equations the variance among the X_{ij} in the population has been assumed to be σ^2, regardless of family distribution. The correction to σ^2 necessary to take account of selection among parents becomes of order h^6 or h^8 in equation (2) and contributes little. In addition, we have ignored the effects, which may be more serious, of selection on the variance in later generations. We shall use the corrected phenotypic variances, such as $(1 - \frac{1}{2}h^4)\sigma^2/M$ solely to permit a more direct comparison with other heritability estimation procedures which use the regression of offspring on parent.

Collecting terms in (3), we have for a Poisson family size distribution,

$$\begin{aligned} V(\bar{X}_{t+1} - \bar{X}_1 \mid \bar{Y}_1 - \bar{X}_1 , \cdots , \bar{Y}_t - \bar{X}_t) \\ &= (\sigma^2/N)\{p(1 - h^2)^2 + h^2(1 - h^2) \\ &\quad + (t - 1)h^2[1 - h^2(1 - p)] + p(1 - \tfrac{1}{2}h^4)\} \\ &= (\sigma^2/N)\{th^2[1 - h^2(1 - p)] + p(2 - \tfrac{1}{2}h^2) - \tfrac{1}{2}ph^4\}. \end{aligned} \qquad (4)$$

If the numbers recorded differ from generation to generation, the appropriate values must be inserted in the separate parts of (4).

In the above derivation no variance due to fluctuations in the environment common to all individuals in a generation is included. Let these effects be independently distributed in successive generations, with mean 0 and variance σ_e^2. Then $V(\bar{X}_{t+1} - \bar{X}_1 \mid \bar{Y}_1 - \bar{X}_1 , \cdots , \bar{Y}_t - \bar{X}_t)$, given by (4), is increased by $2\sigma_e^2$. The variances and covariances of the observed generation means, in each case conditional on the appropriate observed selection differentials, are as follows, using the approximate formulae for Poisson family sizes:

$$V(\bar{X}_t) = \sum_{i=1}^{t-1} h^2[1 - h^2(1 - p_i)]\sigma^2/N_i + (1 - \tfrac{1}{2}h^4)\sigma^2/M_t + \sigma_e^2 \qquad t > 1,$$

$$\operatorname{cov}(\bar{X}_t , \bar{X}_{t'}) = \sum_{i=1}^{t-1} h^2[1 - h^2(1 - p_i)]\sigma^2/N_i + h^2\sigma^2/M_t \qquad t < t',$$

where $p_i = N_i/M_i$, and N_i and M_i are the numbers selected and recorded in generation i. The covariance of generation means comprises the drift variance present in the earlier, t, of the two generations, plus a term $h^2\sigma^2/M_t$ deriving from the covariance of the observed selection differential and the estimate of the mean at generation t. In all later formulae we shall assume that N_i and M_i remain constant each generation.

(iii) *Divergent selection*

In a scheme of divergent selection from the initial population the highest and lowest ranking individuals are selected. In subsequent generations two separate populations are maintained, selected in opposite directions, and the mean performance of the two compared. With animals reared at the same time, common environment effects are thereby eliminated. Let R_i be the difference in observed means after i selections, and let s_i be the sum of the up and down selection differentials obtained at generation i. Thus, if

$\bar{X}_{ui}$, $\bar{X}_{di}$, etc., represent the appropriate means for up and down selection, then

$$R_i = \bar{X}_{u,i+1} - \bar{X}_{d,i+1} ,$$

$$s_i = (\bar{Y}_{ui} - \bar{X}_{ui}) - (\bar{Y}_{di} - \bar{X}_{di}), \; i > 1, \quad \text{and} \quad s_1 = \bar{Y}_{u1} - \bar{Y}_{d1} .$$

In the first generation, errors of estimation of the selection differentials are seen to cancel if the same initial population is used. Letting $S_t = \sum_{i=1}^{t} s_i$, we have

$$E(R_t \mid s_1 , \cdots , s_t) = h^2 S_t \tag{5}$$

and, for a Poisson family size distribution

$$V(R_t \mid s_1 , \cdots , s_t)$$

$$= \frac{2\sigma^2}{N} \{h^2(1 - h^2) + (t - 1)h^2[1 - h^2(1 - p)] + p(1 - \tfrac{1}{2}h^4)\}$$

$$= \frac{2\sigma^2}{N} \{th^2[1 - h^2(1 - p)] + (1 - \tfrac{3}{2}h^4)p\} \tag{6}$$

In later formulae the conditioning on the s_i will be implied and we shall write simply $V(R_t)$.

It may be possible to replicate the selection experiment, where the total of M recorded individuals are split into n replicated subpopulations each with M/n recorded, and a proportion p in each selected every generation. From equation (6) we see that the variance of response between replicates is

$$(2n\sigma^2/N)\{th^2[1 - h^2(1 - p)] + (1 - \tfrac{3}{2}h^4)p\}.$$

The variance of response of each subpopulation is n times as large, and the variance of the mean response over all subpopulations is equal to that of the single population of size M, so that replication does not improve the estimate of response or realised heritability if the same total facilities are used. However, in a replicated experiment the variance of response can be estimated directly from the variance among replicates, rather than from the approximate predictive formulae developed here.

We are primarily concerned with dioecious populations, and our formulae can readily be generalised. Let M_m and N_m be the number of males recorded and selected, respectively, with $p_m = N_m/M_m$, and let M_f , N_f , and p_f be the corresponding parameters for females. With divergent selection, equation (6) with a Poisson distribution of family sizes becomes

$$V(R_t) = \tfrac{1}{2}\sigma^2\left\{\frac{th^2}{N_m}[1 - (1 - p_m)h^2] + \frac{th^2}{N_f}[1 - (1 - p_f)h^2]\right.$$

$$\left. + \left(\frac{1}{M_m} + \frac{1}{M_f}\right)(1 - \tfrac{1}{2}h^4)\right\}. \tag{7}$$

If the same proportion of males and females are recorded, then (7) reduces to (6) with N replaced by N_e , the effective population size, where $1/N_e = 1/4N_m + 1/4N_f$.

REALISED HERITABILITY ESTIMATION

(i) *Experimental design*

The estimator of realised heritability which we use is $\hat{h}^2 = R_t/S_t$, which has variance $V(\hat{h}^2) = V(R_t)/S_t^2$, and is unbiassed (equation (5)).

With mass selection the expected total selection differential with divergent selection is $S_t = 2ti\sigma$, where i is the standardised selection differential. We shall assume in our prediction formulae that the value of $2ti\sigma$ is actually obtained. For simplicity we consider experiments in which the same number of males and females are recorded, and each sex is selected with the same intensity; this design actually minimises the sampling variance of the heritability estimate for a fixed total number, $M = M_m + M_f$, recorded. From (7)

$$V(\hat{h}^2) = \frac{h^2}{2tMpi^2}\left[1 - h^2(1 - p) + (1 - \tfrac{3}{2}h^4)\frac{p}{th^2}\right]. \tag{8}$$

In the experiment a total of $T = M(2t + 1)$ individuals are recorded, comprising M in the base population and in each of the high and low selected lines in the following t generations. Equation (8) becomes

$$V(\hat{h}^2) = \frac{h^2}{Tpi^2}\left(\frac{2t + 1}{2t}\right)\left[1 - h^2(1 - p) + (1 - \tfrac{3}{2}h^4)\frac{p}{th^2}\right]. \tag{9}$$

There is clearly an optimal value of p which minimises $V(\hat{h}^2)$ for a given value of t and h^2, and Soller and Genizi [1967] discussed this problem using their formulae. Assuming that phenotypes are normally distributed, and that the selected populations are of sufficient size so that $i = z/p$, where z is the ordinate of the standardised normal curve, the optimal proportion selected can readily be found by trial and error. The results are summarised in Table 1 for a range of h^2 and t values, together with the associated sampling variances of the heritability estimate.

If the trait is thought to have a very low heritability, intense selection (about 6%) should be practised if the experiment is to run a short time. In a long-term experiment ($t \to \infty$) with low h^2 ($\to 0$), the terms $(2t + 1)/2t$ and $1 - h^2(1 - p) + (1 - \frac{3}{2}h^4)p/th^2$ tend to unity, and equation (9) reduces to $V(\hat{h}^2) = h^2/Tpi^2$, which is minimised when $p = 0.27$. This optimal intensity of selection among the parents also applies to a one-stage selection ($t = 1$) if many more progeny than parents are recorded (Soller and Genizi [1967]).

For traits of high heritability the optimal intensity of selection does not depend greatly on t, since the term, $(1 - \frac{3}{2}h^4)p/th^2$, in equation (9) which includes the error of estimation of the genetic mean in the last generation is small relative to the drift term, $1 - h^2(1 - p)$. Further, $V(\hat{h}^2)$ is influenced very little by the length of the experiment if h^2 is high, and most of the small reduction with increasing t is included in the expression $(2t + 1)/2t$, which can be viewed as the loss of efficiency associated with having one more generation of recording than selection. Even with traits of intermediate to

TABLE 1

OPTIMAL PROPORTION SELECTED, p, AND SAMPLING VARIANCE, $v = 100TV(\hat{h}^2)$, FOR A REALISED HERITABILITY ESTIMATE WITH T RECORDED INDIVIDUALS. THESE ARE COMPARED WITH THE OPTIMAL DESIGN FOR ESTIMATION BY OFFSPRING-PARENT, O–P, REGRESSION WITH ASSORTATIVE MATING (n = OPTIMAL FAMILY SIZE)

h^2	.05		.1		.2		.4		.6		.8	
t	p	v	p	v	p	v	p	v	p	v	p	v
1	.06	68	.09	92	.12	125	.15	156	.15	150	.14	108
2	.09	40	.12	56	.16	82	.17	110	.17	113	.14	89
5	.14	24	.17	36	.20	58	.20	85	.18	92	.14	80
10	.17	18	.20	30	.22	50	.22	77	.18	86	.14	74
20	.20	15	.22	26	.23	46	.22	73	.19	83	.14	72
$\rightarrow\infty$	.26	12	.26	23	.26	42	.22	69	.20	80	.14	70
O–P	.06	64	.07	82	.10	104	.11	121	.12	114	.11	85
n	27		17		11		8		7		8	

low heritability, equation (9) shows that after a few generations most of the sampling variance of the estimate is contributed by genetic drift, and we see in Table 1 that continuing the experiment for many generations improves the heritability estimate little when the total number recorded is fixed. Of course, if the only criterion of cost is the number measured per generation, efficiency is improved by increasing the length of the experiment, and if most of the variance comes from genetic drift, $V(\hat{h}^2) \propto 1/t$, approximately.

In a realised heritability estimate the regression of offspring on parent is calculated from the mean performance of all progeny and the mean performance of all selected parents. In other methods the regression is computed among individual families, so that the variation between families is utilised, and only a single generation of parents and their progeny are recorded. The most efficient design for a fixed total number of parents and progeny recorded is obtained if only the best and poorest potential parents are selected, and these are mated assortatively. The optimal proportion which should be selected, and the associated optimal progeny family size have been obtained by Hill [1970]. The essential results are included in Table 1 so that the efficiencies of the offspring-parent (O–P) method and realised heritability methods can be compared with the same total number, T, recorded. The O–P method is rather more efficient than a single generation selection experiment with high heritability values, but the difference is very small if heritability is low. However, a realised heritability estimate from a selection experiment of more than 3 generations always has a lower sampling variance than an estimate from offspring-parent regression.

There are three factors which contribute to the higher sampling variances in a single generation with the realised heritability method relative to the O–P method, especially at high h^2 values. These are: (1) family sizes are

assumed to be fixed in the O–P method. Imposing the same restriction for realised heritability estimation, the term $1 - \frac{3}{2}h^4$ in equation (9) becomes $1 - \frac{1}{2}h^2 - h^4$. (2) Variation between parent family means, $[1 - i(i - x)]\sigma^2$, within the high and low selected groups is used in the O–P method to increase the sum of squares for the independent variate, where x is the abscissa of the standardised normal curve associated with p. The total variance is then $[i^2 + 1 - i(i - x)]\sigma^2 = (1 + ix)\sigma^2$, and only differs greatly from i^2 with the weaker selection optimal at high h^2 values (Table 1). (3) On average, families in the realised heritability method comprise $1/p$ males and $1/p$ females, to give $2/p$ in all. In the O–P method the family sizes and proportions selected can be manipulated separately. Departures of the optimal family size from $2/p$ for the O–P method in Table 1 are relatively larger at high h^2 values.

The design of a selection experiment which minimises the sampling variance of a realised heritability estimate depends on the heritability value itself, but this is an unknown parameter. Therefore it is important for us to know both how far the proportion selected can depart from the optimal values given in Table 1 before efficiency is much affected, and also what designs are robust against poor prediction of the parameter. Soller and Genizi [1967] discuss this problem with their model, and find that p values in the range 0.15–0.20 generally give satisfactory results, but they consider primarily a value of 0.5 for h^2. Some results with our model are summarised in Table 2, in which there are three examples with optimal p values of 0.12, 0.20, and 0.26. Although these results were obtained using only $h^2 = 0.2$, it is clear from equation (9) that the relative efficiency of estimation at p values away from the optimum is solely a function of the optimal p, and not of h^2 or t. Therefore the results in Table 2 for $h^2 = 0.2$, $t = 5$ are approximately those which would be obtained for $h^2 = 0.05$, $t = 20$, since $p = 0.2$ is the best design in both cases. We see in the table that the efficiency of estimation is rather insensitive to changes in the proportion selected. For example, if $p = 0.2$ at the optimum, an efficiency of over 90% can be obtained by selecting anywhere between 0.11 and 0.32 each generation. In agreement with Soller and Genizi, we find that p values of 0.15–0.20 are generally efficient.

TABLE 2

EFFECTS OF DEPARTURE FROM OPTIMAL PROPORTION SELECTED, p_0, IN REALISED HERITABILITY ESTIMATION FOR $h^2 = 0.2$ AND DURATION t GENERATIONS. TABLE ENTRIES ARE $100[V(\hat{h}^2)$ WITH p_0 SELECTED$]/[V(\hat{h}^2)$ WITH p SELECTED$]$

t	p_0	p .04	.08	.12	.16	.20	.24	.28	.32	.36	.40
1	.12	78	96	100	97	92	85	77	70	63	56
5	.20	58	81	93	99	100	98	95	90	84	78
$\rightarrow\infty$	.26	48	71	85	94	98	100	99	97	94	89

(ii) *Comparison with other methods*

We can use our results to compare the sampling variances of heritability estimates obtained from selection experiments, or other regression designs, with those obtained from the covariance of full or half sibs. Robertson [1959a] showed that with an optimal design $V(\hat{h}^2) = 16h^2(1 - \frac{1}{2}h^2)^2/T$, approximately, from the covariance of full sibs, and $32h^2(1 - \frac{1}{4}h^2)^2/T$, approximately, from the covariance of half sibs.

Since sib covariance designs can be used for traits which can only be measured on one sex, we need to develop our realised heritability theory to include this situation. Imagine that the population comprises pair matings, with individuals of one sex chosen by mass selection, and those of the other unrecorded and chosen at random. Divergent selection is practised, with a total number T recorded in an experiment of t generations. It can be shown, with a Poisson distribution of family sizes, that

$$V(\hat{h}^2) = \frac{2h^2}{Tpi^2}\left(\frac{2t+1}{2t}\right)\left[1 - \tfrac{1}{2}(1-p)h^2 + (2-h^4)\frac{p}{th^2}\right],$$

which can be compared directly with equation (9). For long-term experiments with low h^2, the efficiency with one sex recorded is one-half that where both sexes are recorded. With shorter term experiments, or higher heritabilities, there is a greater difference in efficiency.

Using these formulae, and other results from Hill [1970] for offspring-parent regression with unselected parents, or with measurements on only one sex, various methods of estimating heritability are compared in Table 3. In each case the appropriate optimal design is used. Realised heritability

TABLE 3

COMPARISON OF ALTERNATIVE METHODS OF ESTIMATING HERITABILITY. RESULTS ARE EXPRESSED AS $v = 100TV(\hat{h}^2)$, FOR A TOTAL OF T INDIVIDUALS RECORDED, (O–P = OFFSPRING-PARENT REGRESSION)

	Sexes recorded	h^2 .1	.4
Realised heritability (t = 5 generations)	2	36	85
O–P, selection and assortative mating	2	82	121
O–P, no selection, assortative mating	2	163	192
O–P, no selection, no assortative mating	2	325	384
Realised heritability (t = 5 generations)	1	94	230
O–P, selection	1	252	394
O–P, no selection	1	570	672
Covariance of full sibs	1 or 2	144	410
Covariance of half sibs	1 or 2	304	1037

estimates are clearly best, but of course take much longer to obtain. If selection among parents is practised, standard offspring-parent regression techniques, taking a total of only two generations, compare very favourably with the full or half-sib covariance methods. The conclusion of Robertson [1959a] and Falconer [1960] that regression methods are less efficient at heritabilities below 20–25% holds only if unselected parents are used.

UTILISATION OF INFORMATION FROM RELATIVES

In our earlier discussion we have assumed that individuals are selected solely on their own performance for some quantitative character, yet in many programmes measurements on the same trait on relatives are combined into a selection index. In the simplest case the index may be only the full sib family mean. Unless mass selection is practised, realised heritability estimation is more difficult, in that the ratio of response to selection differential depends on h^2. However, an index may be used in comparisons of alternative selection schemes, or to detect whether genetic variation exists in a population. In such situations, a useful criterion of efficiency of a scheme is the ratio $E^2(R_t)/V(R_t)$, which is also the inverse of the square of the coefficient of variation of response.

Imagine individuals are selected on an index with a correlation r_{IG} between the index and breeding value. With selection of equal intensity in each sex, divergent selection, and Poisson family distribution, it can be shown that

$$V(R_t) = (2\sigma^2/N_e)\{th^2[1 - r_{IG}^2(1 - p)] + (1 - \tfrac{3}{2}h^2r_{IG}^2)p\} \qquad (10)$$

and

$$E(R_t) = 2tihr_{IG}\sigma,$$

where N_e is the effective population size. With mass selection $r_{IG} = h$ and equation (10) reduces to our earlier formulae. Generalisation of (10) to other situations, such as use of a different index in the two sexes, is straightforward.

If we consider experiments of, say, 5 or more generations, the term from drift variance, $th^2[1 - r_{IG}^2(1 - p)]$ in (10) predominates, and we have

$$E^2(R_t)/V(R_t) = 2N_eti^2r_{IG}^2/[1 - r_{IG}^2(1 - p)]$$

approximately, and

$$E^2(R_t)/V(R_t) \propto N_er_{IG}^2/[1 - r_{IG}^2(1 - p)]$$

approximately, if t and i are the same in alternative schemes.

Let us consider an idealised population under selection in which M individuals are recorded and Mp are selected each generation. There is a pair mating structure in which each of the $\frac{1}{2}Mp$ families have $1/p$ male and $1/p$ female progeny, where $1/p$ is integral, and family means are estimated from the mean of both sexes. The phenotypic intra-class correlation of family members is k, where $k = \frac{1}{2}h^2$ if there are no common environment effects of family members. We consider three alternative selection criteria: mass

selection, selection within families where the index is the deviation of the individual observation from the family mean, and between family selection where the index is the family mean itself, such that a proportion p of entire families is chosen. The effective population size for selection within families is double that for random family size (Falconer [1960]); and between family selection is equivalent to picking all members of Np of the families in the previous generation, so the effective size is $2Mp^2$. The correlations r_{IG} depend on the number measured in each family, and thus on p in our model. The necessary formulae are summarised below.

	N_e	r^2_{IG}
Mass selection	Mp	h^2
Within family selection	$2Mp$	$\dfrac{(2-p)h^2}{8(1-k)}$
Between family selection	$2Mp^2$	$\dfrac{(2+p)^2h^2}{8[p+(2-p)k]}$

Values of $E^2(R_t)/V(R_t)$ are compared for these three schemes in Table 4, where low values in the table indicate inefficient designs. In these examples, with comparisons made at the same value of p, we see that within family selection is never as efficient as mass selection, even when there are quite large common environmental effects of family members, because the increased population size does not compensate for the reduced response. Between

TABLE 4

SENSITIVITY OF ALTERNATIVE SELECTION SCHEMES, EXPRESSED AS $[100E^2(R_t)/Mh^2V(R_t)]$, FOR SPECIFIED p = PROPORTION SELECTED, h^2 = HERITABILITY AND k = PHENOTYPIC INTRACLASS CORRELATION OF FULL SIBS. THE SENSITIVITY IS ALSO GIVEN AT p_0, THE OPTIMAL VALUE OF p, TOGETHER WITH p_0(%)

Method of selection	p	h^2 = $\frac{1}{8}$, k = $\frac{1}{16}$	h^2 = $\frac{1}{8}$, k = $\frac{1}{8}$	h^2 = $\frac{1}{4}$, k = $\frac{1}{8}$	h^2 = $\frac{1}{4}$, k = $\frac{1}{4}$	h^2 = $\frac{1}{2}$, k = $\frac{1}{8}$	h^2 = $\frac{1}{2}$, k = $\frac{1}{2}$
Mass	$\frac{1}{8}$	381	381	434	434	603	603
	$\frac{1}{4}$	446	446	497	497	646	646
	$\frac{1}{2}$	339	339	364	364	425	425
	p_0	446(26)	446(26)	497(25)	497(25)	653(21)	653(21)
Within family	$\frac{1}{8}$	174	187	193	227	245	400
	$\frac{1}{4}$	193	207	212	249	265	423
	$\frac{1}{2}$	129	125	140	164	170	263
	p_0	194(22)	208(22)	213(22)	251(22)	267(22)	430(22)
Between family	$\frac{1}{8}$	265	161	203	102	138	59
	$\frac{1}{4}$	426	312	365	225	284	144
	$\frac{1}{2}$	456	390	422	320	366	236
	p_0	482(39)	397(45)	437(41)	321(46)	373(45)	236(51)

family selection is only more efficient than mass selection when the heritability is low, there is little common environmental variance, and selection is not intense. With mass selection the effective population size may be reduced by selection to a value less than Mp (Robertson [1961]). The results may therefore be somewhat biassed in favour of mass selection. Nevertheless mass selection appears the most satisfactory design for comparing alternative schemes or responses from different populations or different traits.

However, the optimal intensity of selection for maximising efficiency is not the same for each selection scheme. Table 4 also contains values of $E^2(R_t)/V(R_t)$ at the value of p at which this is maximised, together with the appropriate p. For these, essentially long term experiments since measurement error is ignored, we find that at low heritabilities p should be around 0.26 for mass selection, 0.22 for selection within families and 0.45 for between family selection. Even when compared at their optimal p values, mass selection is usually most efficient.

CORRELATED TRAITS

(i) *Realised genetic correlations*

Selection experiments can also be used for estimation of genetic correlations by selecting for one trait, say X, in one population and the other trait, Y, in a second population and observing direct and correlated responses (Falconer [1960]). We denote the direct response in X when selecting for X as R_X , and the associated correlated response in Y as C_Y , both measured in the final generation, and we denote R_Y and C_X similarly. We assume in this discussion that divergent selection is practised, so there are 4 populations in all.

Falconer [1960] gives the following estimator of the realised genetic correlation

$$\hat{r} = (C_X C_Y / R_X R_Y)^{\frac{1}{2}}. \tag{11}$$

Since R_X and C_Y are measured in one population, and R_Y and C_X in another, the ratio C_Y/R_X is uncorrelated with the ratio C_X/R_Y . If we define $\gamma^2_{R_X} = V(R_X)/E^2(R_X)$ and $\gamma_{R_X C_Y} = \text{cov}\,(R_X\ ,\ C_Y)/E(R_X)E(C_Y)$, for example, and assume these coefficients of variation are small, we obtain from (11)

$$V(\hat{r}^2) = r^4(\gamma^2_{R_X} - 2\gamma_{R_X C_Y} + \gamma^2_{C_Y} + \gamma^2_{R_Y} - 2\gamma_{R_Y C_X} + \gamma^2_{R_X}) \tag{12}$$

approximately. We can show that, for example,

$$V(R_X) = (2\sigma_X^2/N_*)\{th_X^2[1 - h_X^2(1-p)] + (1 - \tfrac{3}{2}h_X^4)p\},$$

$$V(C_Y) = (2\sigma_Y^2/N_*)\{th_Y^2[1 - h_X^2 r^2(1-p)] + (1 - \tfrac{3}{2}h_X^2 h_Y^2 r^2)p\},$$

$$\text{cov}\,(R_X\ ,\ C_Y) = (2\sigma_X\sigma_Y/N_*)\{th_X h_Y r[1 - h_X^2(1-p)] + (r_P - \tfrac{3}{2}h_X^3 h_Y r)p\},$$

$$E(R_X) = tih_X^2\sigma_X\ , \qquad E(C_Y) = tih_X h_Y r\sigma_Y\ ,$$

where the same proportion, p, is selected in each line and sex, with Poisson family sizes. Also σ_X^2 , σ_Y^2 , h_X^2 , h_Y^2 are the appropriate phenotypic variances and heritabilities, r is the genetic correlation, and r_P the phenotypic correlation between X and Y. Substituting into (12) we obtain

$$V(\hat{r}^2) = \frac{r^2}{2N_e t i^2 h_X^2 h_Y^2}\left\{(1-r^2)(h_X^2+h_Y^2) + \frac{p}{t}\left[\frac{1}{h_Y^2}(r^2h_X^2 - 2r_P h_X h_Y + h_Y^2) + \frac{1}{h_X^2}(r^2h_Y^2 - 2r_P h_X h_Y + h_X^2)\right]\right\} \quad (13)$$

approximately. If $\hat{r}$ has a small coefficient of variation, $V(\hat{r}) = V(\hat{r}^2)/4r^2$, and if t is large, so that we need only consider the drift terms in (13),

$$V(\hat{r}) = \frac{1-r^2}{8N_e t i^2}\left(\frac{1}{h_X^2} + \frac{1}{h_Y^2}\right), \quad \text{approximately,} \quad (14)$$

If $h_X^2 = h_Y^2$ and $r_P = r$, equation (13) becomes

$$V(\hat{r}) = \frac{1}{4N_e t i^2 h^2}\left[1 - r^2 + \frac{(1-r)^2 p}{th^2}\right], \quad \text{approximately.}$$

When the heritabilities of X and Y are estimated in the same experiment, equation (15) may be written

$$V(\hat{r}) = \frac{1-r^2}{4h_X^2 h_Y^2}[V(\hat{h}_X^2) + V(\hat{h}_Y^2)], \quad \text{approximately.} \quad (15)$$

where we have ignored terms in $h^2(1-p)$ relative to 1 in $V(\hat{h}^2)$. In this arrangement, however, (15) resembles (especially when $h_X^2 = h_Y^2$) the formula of Robertson [1959b] which can be applied to other methods of heritability and genetic correlation estimation, namely

$$V(\hat{r}) = \frac{(1-r^2)^2}{2h_X^2 h_Y^2}[V(\hat{h}_X^2)V(\hat{h}_Y^2)]^{\frac{1}{2}}.$$

Most important, equation (15) indicates that designs which are optimal for estimation of realised heritabilities will also be satisfactory for estimation of realised genetic correlations; and since selection experiments are found to be efficient systems for estimating heritabilities they must be efficient for estimating genetic correlations also.

(ii) *Marker genes*

A possible method for estimating the effect, if any, on a quantitative trait of genes at a marker locus at which the individual genotypes can be identified, is to select the trait in a population in which the marker is segregating, and observe changes in its frequency. We need some measurement of the sensitivity of such experiments and information on optimal design. Let us assume divergent selection is practised, that each of the three genotypes of the marker can be identified, and that it has an additive effect on the trait X, with a difference, a, between the homozygotes in effect. We let the

gene frequency be q, and regard q as a new trait and cast the results in our earlier framework. Thus

$$\sigma_q^2 = \tfrac{1}{2}q(1-q), \qquad h_q^2 = 1, \quad \text{and} \quad r_{Xq}^2 = a^2q(1-q)/2h_X^2\sigma_X^2 .$$

We can also view r_{Xq}^2 as the proportion of the additive variance contributed by the marker locus. If we select on an index I, of X (e.g. family selection), then assuming normality, $r_{Iq} = r_{IX}r_{Xq}$, and

$$E^2(R_q) = 2t^2i^2r_{IX}^2r_{Xq}^2q(1-q),$$

$$V(R_q) = \frac{q(1-q)}{N_e}\{t[1 - r_{IX}^2r_{Xq}^2(1-p)] + (1 - \tfrac{3}{2}r_{IX}^2r_{Xq}^2)p\}.$$

Assuming r_{Xq}^2 is small, and that selection proceeds for several generations,

$$E^2(R_q)/V(R_q) = 2tN_ei^2r_{IX}^2r_{Xq}^2$$

is a good approximation.

Our earlier comparisons of the efficiency of alternative index designs therefore hold approximately. To estimate effects of marker genes the optimal proportion selected is 0.27 for mass selection, 0.22 for within family selection, and about 0.45 (0.42 for $h^2 = \frac{1}{8}$ to 0.5 for $h^2 = \frac{1}{2}$) for between family selection if there are no common environment effects of family members. At the optimal proportions, within family selection is from 47% to 58% and between family selection from 106% to 70% as efficient as mass selection for h^2 ranging from $\frac{1}{8}$ to $\frac{1}{2}$, again assuming no family environment effects. Mass selection can therefore be recommended for general application.

With mass selection, the estimator of a/σ (the gene effect as a proportion of the phenotypic standard deviation) is $R_q/[tiq(1-q)]$, which has variance $1/[N_etiq(1-q)]$.

(iii) *A practical problem*

Our results for the standard errors of realised genetic correlations are very approximate, requiring small coefficients of variation of the direct and correlated responses. Frequently our objective in an experiment is not merely to estimate heritabilities or correlations, but to ask a more specific question, such as 'which scheme gives more rapid progress?' or 'is there genetic variance in some trait?' A worked example of the design required to solve a practical problem (which, in fact, initiated this whole investigation) may help as illustration.

Imagine we wish to improve pigs for food conversion efficiency under *ad lib.* or restricted feeding, and need to know whether there is an interaction of genotype with these environments. Let us assume that the interaction is unimportant if the genetic correlation of food conversion efficiency on the two rations is greater than 0.8. Alternatively, if we select on one diet and observe the correlated response on the other we wish to know whether the indirect response is at least 80% of the direct response. The experiment could comprise two lines started from the same base population, with mass

selection practised in each sex and an idealised family structure of 1 boar to 4 sows, with 4 male and 4 female progeny recorded in each litter. The proportion of males selected is $\frac{1}{16}$, and of females $\frac{1}{4}$, giving $\bar{\imath} = 1.6$, and $N_e = 3.2N_m$. We assume that for each trait $h^2 = 0.4$. If the experiment is continued, say 4 or more generations, we can ignore the measurement error for the final generation in equation (7). Under the null hypothesis that $r = 1$, we obtain from (7),

$$V(R_X - C_X) = \tfrac{1}{2}\sigma_X^2 h_X^2 t\left\{\frac{1}{N_m}[1 - (1 - p_m)h_X^2] + \frac{1}{N_f}[1 - (1 - p_f)h_X^2]\right\}$$

$$= 0.225t\sigma_X^2/N_m .$$

Under the null hypothesis, $E(R_X - C_X) = 0$, and under the alternative hypothesis ($r = 0.8$)

$$E(R_X - C_X) = tih_X^2\sigma_X(1 - rh_Y/h_X) = 0.128t\sigma_X .$$

Therefore

$$E(R_X - C_X)/[V(R_X - C_X)]^{\frac{1}{2}} = 0.270\ (tN_m)^{\frac{1}{2}}.$$

For one-tail tests with 5% type 1 error and 80% power we require that the ratio of difference to its standard error exceeds 2.5, approximately. Therefore we need $tN_m > 86$, or for an experiment of 4 generations, N_m (the number of selected males) ≥ 22 on each treatment. A large experiment is clearly required.

DISCUSSION

In our model we have excluded environmental effects common to all individuals in the population in any generation by utilising designs involving divergent selection, where these effects are eliminated if the lines are maintained contemporaneously in the same environment. Alternatively, we could compare response with that in an unselected control population, for which the mean would have variance in generation t of $\sigma^2(th^2 + p)/N_e$, where p is the ratio of its effective size, N_e, to the total number recorded. But more efficient ($\times$ 2, approximately) estimates of realised heritability or genetic correlation will be obtained from the same facilities using divergent selection rather than a control population. It has been shown that if common environmental effects are assumed to be randomly independently distributed with constant mean and variance σ_c^2 over generations, the variance of total response is inflated by $2\sigma_c^2$ if no control or comparable selection line is maintained. Should there be a trend in the environment, estimators of response will have an unknown bias. Even with a control population, or with comparison of selection lines, the sampling variance could be inflated by genotype-environment interaction, should the two populations react differently to the changes in common environment during the experiment. Fortunately such interaction seems unlikely to be important in short-term experiments if the selection lines or control all originate from the same base population.

A basic assumption of the model needs to be stressed, namely that the genetic and environmental variances and covariances remain constant in each population during the selection programme. If there is much inbreeding, or there are genes with a large effect on the quantitative trait under selection, changes in the genetic variance are likely to occur. Therefore our results are probably of most relevance to experiments of only a few generations duration. Even then we find that most of the sampling variance of a selection response or realised heritability estimate is contributed by genetic drift, rather than by inaccurate estimation of the final genetic mean from recording only a finite number of individuals. Consequently there is little advantage in measuring large numbers in the last generation, and we find that the accuracy of a realised heritability estimate is largely a function of the total number recorded and selected over the whole experiment, rather than its duration, or the numbers measured per generation, taken separately.

ACKNOWLEDGMENTS

The author is indebted to Professors Alan Robertson and O. Kempthorne for their helpful comments and suggestions.

PLANIFICATION ET EFFICACITE DES EXPERIENCES DE SELECTION POUR ESTIMER LES PARAMETRES GENETIQUES

RESUME

(1) On établit des formules pour la variance d'échantillonnage de la réponse à la sélection et pour les estimations de l'héritabilité obtenue et de la correlation génétique obtenue. (2) Si l'on garde une population témoin, ou si l'on pratique une sélection divergente, la plus grande partie de la variance d'échantillonnage provient de la dérive génétique et dépend d'avantage du nombre total d'individus dénombrés dans la totalité de l'expérience que de la durée de l'expérience. (3) On recherche l'intensité de sélection optimale pour estimer les héritabilités obtenues, des proportions de 15 pour cent environ de selection seraient satisfaisantes. Des plans du même type se montreraient aussi efficaces pour l'estimation des correlations génétiques obtenues. (4) On compare plusieurs méthodes d'estimation de l'héritabilité, de celles-ci l'héritabilité obtenue est celle qui a la plus petite variance. (5) On évalue quelques indices de sélection pour améliorer un seul caractère. La sélection de masse est sans doute la meilleure méthode de comparaison des réponses de programmes de sélection alternative ou de populations.

REFERENCES

Basu, D. [1955]. On statistics independent of a complete sufficient statistic. *Sankhyā 15*, 377–80.

Cochran, W. G. [1951]. Improvement by means of selection. *Proc. Second Berkeley Symp. Math. Stat. Prob.* pp. 449–70.

Falconer, D. S. [1960]. *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh.

Finney, D. J. [1956]. The consequences of selection for a variate subject to errors of measurement. *Bull. Inst. Int. Statist. 24*, 1–10.

Finney, D. J. [1961]. The transformation of a distribution under selection. *Sankhyā A23*, 309–24.

Hill, W. G. [1970]. Design of experiments to estimate heritability by regression of offspring on selected parents. *Biometrics 26*, 566–71.

Pearson, K. [1903]. On the influence of natural selection on the variability and correlation of organs. *Phil Trans. R. Soc. Lond. A200*, 1–66.

Prout, T. [1962]. The error variance of the heritability estimate obtained from selection response. *Biometrics 18*, 404–7.

Richardson, R. M., Kojima, K., and Lucas, H. L. [1968]. An analysis of short term selection experiments. *Heredity 23*, 493–506.

Robertson, A. [1959a]. Experimental design in the evaluation of genetic parameters. *Biometrics 15*, 219–26.

Robertson, A. [1959b]. The sampling variance of the genetic correlation coefficient. *Biometrics 15*, 469–85.

Robertson, A. [1961]. Inbreeding in artificial selection programmes. *Genet. Res. 2*, 189–94.

Soller, M. and Genizi, A. [1967]. Optimum experimental designs for realised heritability estimates. *Biometrics 23*, 361–65.

Received April 1970, *Revised November* 1970

11

Estimation of realised heritabilities from selection experiments

I. Divergent selection

by

William G. Hill

Reprinted from
BIOMETRICS *28*, 747–765
September 1972

ESTIMATION OF REALISED HERITABILITIES FROM SELECTION EXPERIMENTS I. DIVERGENT SELECTION

WILLIAM G. HILL

Institute of Animal Genetics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN Scotland.

SUMMARY

Methods of estimating realised heritability from selection experiments are compared. For designs in which divergent selection is practiced, formulae are given for the sampling variance of some simple linear estimators of realised heritability, such as the regression of cumulative response on cumulative selection differential. Although the variance-covariance structure of the responses depends on the heritability, it is found that for most relevant combinations of parameters these linear estimators are almost as efficient as a maximum likelihood (ML) estimator, and can be recommended for practical use. Standard methods of calculating the variance of these estimators are shown to be very biased, downwards for the regression of cumulative response on cumulative selection differential. Methods of estimating the variance from experimental data, which are almost unbiased, are described.

1. INTRODUCTION

The heritability of a quantitative trait can be estimated from the regression of response on selection differential in selection experiments or breeding programmes continued for a few generations. Falconer [1954; 1960] has called such estimates 'realised heritability estimates', since they describe the results of the selection. The realised heritabilities can be used to check predictions made prior to the experiment, or alternatively to provide estimates of heritability which are more precise than can be obtained by other methods. Aspects of the design and efficiency of experiments for realised heritability estimation have been discussed in an earlier paper (Hill [1971]), in which a simple estimator was used, the ratio of total response to total selection differential. More commonly, the regression of the cumulative response on cumulative selection differential each generation is used as the estimator (Falconer [1960], Richardson *et al.* [1968]). From the error structure derived by Hill [1971] it is clear that neither method is optimal, and it will be shown that the latter method can lead to very biased values for the sampling variance of the realised heritability estimate.

In this paper we shall investigate the sampling variance of a ML estimator of realised heritability. Although this can give more efficient estimates than those obtained by simple linear regression methods, it is biased under some

situations and we shall find that the differences in efficiency are generally small, so that the simple methods can be adopted in practice. Approximate methods for finding the sampling variance of these estimators are discussed. Many assumptions have to be made in defining the model and finding the variance-covariance structure of the observed selection responses. These have been discussed in some detail in the earlier paper Hill [1971] and only trivial further simplifications are adopted here. The most important assumption is that the genetic and phenotypic variances do not change during the course of the experiment. Thus the results are essentially limited to short term experiments.

Several designs of a selection experiment can be distinguished: selection may be practiced in a single direction without a control population or a control can be maintained so that environmental effects common to all individuals in the selected line are removed by measuring the difference between selected and control line means. In an alternative design, divergent selection, the selection is practiced in two lines in opposite directions for the same trait and the difference in performance between the lines recorded. Common environment effects are again eliminated, and more precise estimates of heritability are obtained than with unidirectional selection for the same total facilities, since the sum of squares for the 'independent variable', the selection differential, is increased. In this paper we shall consider only the divergent selection case; we defer the problems of estimating common environmental effects to Part II of this study.

2. MODEL

In the base population, at generation 1, M individuals are recorded and the highest ranking and lowest ranking N selected to start the up-selected and down-selected lines, which are subsequently continued separately, although the generations are contemporaneous. In each generation in both of the lines M individuals are recorded, and the appropriate N selected. The proportion selected is $p = N/M$. We assume for most of the analysis that individuals are monecious, and the minor modifications to the theory required to handle the usual situation of the two sexes are deferred to section 6. The mean performances of the M recorded individuals at generation i are $\bar{X}_{ui}$ and $\bar{X}_{di}$, $i = 1, \cdots, t+1$, and the means of the N selected individuals are $\bar{Y}_{ui}$ and $\bar{Y}_{di}$, $i = 1, \cdots, t$, in the up and down selected lines, respectively. Thus t selections are practiced in all, and the total selection applied in generation i is

$$s_i = (\bar{Y}_{ui} - \bar{X}_{ui}) - (\bar{Y}_{di} - \bar{X}_{di}) \qquad i = 1, \cdots, t,$$

and the cumulative differential up to generation i is given by $S_i = \sum_{j=1}^{i} s_j$. The total response for the first i selections is

$$R_i = \bar{X}_{u,i+1} - \bar{X}_{d,i+1} \qquad i = 0, \cdots, t.$$

We note that $R_0 = 0$, since the same base population is used for both selected lines.

In the trait under selection the phenotypic and genotypic values are assumed to be bivariate normally distributed with means equal to the genotypic mean of the selected individuals of the previous generation, phenotypic variance σ^2, genotypic variance $h^2\sigma^2$, and correlation h, where h^2 is the heritability of the trait. The gene action is assumed to be additive. The error structure is described in detail by Hill [1971], and only the results are summarized here. We have that the responses are multivariate normally distributed, with

$$E(R_i \mid s_1, \cdots, s_t) = h^2 S_i \qquad i = 1, \cdots, t,$$

and the variances and covariances are obtained as follows. The genetic sampling, or drift, variance accumulates over generations and comprises two parts each generation: firstly, $2\sigma^2 h^2(1 - h^2)/N$ for the deviation of the genetic mean of selected individuals about regression and, secondly, $2h^4\sigma^2/M$ for the deviation of the observed selection differentials from their true value, since the observed means $\bar{X}_{wi}$, $\bar{X}_{di}$ deviate by chance from the population mean. In the first generation, when selection is practiced in a single population, this second term is absent since some of the errors cancel. The covariance of pairs of generation means includes these components for the earlier of the two generations, for the selection response is essentially a Markov process. The other component of sampling variance is approximately $2\sigma^2/M$ for the variance of $\bar{X}_{wi} - \bar{X}_{di}$ about its mean and it does not accumulate. In the previous paper a value of $2(1 - \frac{1}{2}h^4)\sigma^2/M$ was derived for this variance, which strictly depends on the mating system; the simplification adopted here has a trivial effect on the results. The following set of variances are all conditional on the observed selection differentials, but the conditioning is omitted from the formulae for brevity. We have

$$\begin{aligned} V(R_i) &= 2\sigma^2[ih^2(1 - h^2)/N + (i - 1)h^4/M + 1/M] \\ &\qquad\qquad i = 1, \cdots, t \\ \operatorname{cov}(R_i, R_j) &= \operatorname{cov}(R_j, R_i) = 2\sigma^2[ih^2(1 - h^2)/N + (i - 1)h^4/M + h^2/M] \\ &\qquad\qquad 1 \leq i < j \leq t \end{aligned} \tag{1}$$

where the last term in cov (R_i, R_j), $2\sigma^2h^2/M$, derives from the covariance of R_i with the observed selection differential for the subsequent generation.

The responses, r_i, in each separate generation will be used in some formulae. They are given by

$$r_i = R_i - R_{i-1} \quad i = 1, \cdots, t$$

where $E(r_i) = h^2 s_i$, and they have a multivariate normal distribution with variance-covariance structure

$$\begin{aligned} V(r_1) &= 2\sigma^2[h^2(1 - h^2)/N + 1/M] \\ V(r_i) &= 2\sigma^2[h^2(1 - h^2)/N + (2 - 2h^2 + h^4)/M] \quad i = 2, \cdots, t \\ \operatorname{cov}(r_i, r_{i+1}) &= \operatorname{cov}(r_{i+1}, r_i) = -2\sigma^2(1 - h^2)/M \quad i = 1, \cdots, t - 1 \\ \operatorname{cov}(r_i, r_j) &= 0 \quad \text{otherwise.} \end{aligned} \tag{2}$$

We shall find it convenient to define a term for the drift variance

$$\sigma_d^2 = 2\sigma^2[h^2(1 - h^2)/N + h^4/M], \tag{3}$$

and a term for the error variance

$$\sigma_e^2 = 2\sigma^2(1 - h^2)/M, \tag{4}$$

such that $V(r_i) = \sigma_d^2 + 2\sigma_e^2 \quad i = 2, \cdots, t,$

$$\text{cov}\,(r_i, r_{i+1}) = \text{cov}\,(r_{i+1}, r_i) = -\sigma_e^2 \quad i = 1, \cdots, t - 1.$$

We also define square matrices of dimension t,

$$\begin{aligned} \mathbf{C} &: c_{ij} = \text{cov}\,(R_i, R_j) \\ \mathbf{P} &: p_{ij} = \text{cov}\,(r_i, r_j) \end{aligned} \tag{5}$$

and column vectors of dimension t,

$$\mathbf{S} = (S_i), \mathbf{s} = (s_i), \mathbf{R} = (R_i), \mathbf{r} = (r_i).$$

3. ESTIMATORS

Simple linear regression theory offers several estimators of realised heritability, which, while not minimum variance except under special conditions, can all be shown to be unbiased. We consider three such linear estimators, together with the ML estimator.

i) *Regression of cumulative response on cumulative selection differential* (b_C)

The regression of cumulative response on cumulative selection differential has been used by Falconer and others, and is defined by

$$b_C = \sum_{i=1}^{t} R_i S_i \Big/ \sum_{i=1}^{t} S_i^2 = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{R}, \tag{6}$$

and

$$E(b_C) = h^2.$$

Under an assumption of $h^2 = 0$, $\mathbf{C}$ is a scalar matrix and b_C is the least squares estimator. In general it has sampling variance

$$V(b_C) = \sum_i \sum_j S_i S_j \text{ cov}\,(R_i, R_j)/(\sum_i S_i^2)^2 = \mathbf{S}'\mathbf{C}\mathbf{S}(\mathbf{S}'\mathbf{S})^{-2}. \tag{7}$$

Some insight into (7) can be obtained for the special case where the selection differentials are constant, even though this could not be achieved in practice. Then we let $s_i = s$ and $S_i = is$ and find that

$$V(b_C) = \frac{6}{s^2 t(t+1)(2t+1)}\left[\frac{2t^2 + 2t + 1}{5}\sigma_d^2 + \sigma_e^2 + \frac{3t(t+1)}{2(2t+1)}h^2\sigma_e^2\right] \tag{8}$$

The last term in (8) arises from the special error structure of the first generation. The term σ_d^2 is of order $1/t$, that in σ_e^2 is of order $1/t^3$ or h^2/t^2, so that after a few generations most variance is contributed by drift unless σ_d^2 is

much smaller than σ_e^2 . When selection differentials are equal, $V(b_C)$ is proportional to the variance of the regression of response (R_i) on generation number. Some formulae similar to (8) have been obtained by Dickerson [1969] for the regression on generations of the mean performance of control populations.

ii) *Regression of individual generation response on selection differential* (b_I)

When the heritability is high (i.e. around 0.5), and selection intense (i.e. N is a small fraction of M) then $\mathbf{P}$ approximates a scalar matrix, except that the first element is smaller. Since $\mathbf{P}$ is the variance-covariance matrix of the r_i , if it is scalar the best linear estimator is the regression of the response on selection differential in individual generations, given by

$$b_I = \sum_{i=1}^{t} r_i s_i \Big/ \sum_{i=1}^{t} s_i^2 = (\mathbf{s}'\mathbf{s})^{-1}\mathbf{s}'\mathbf{r}. \tag{9}$$

This estimator has sampling variance

$$V(b_I) = \sum_i \sum_j s_i s_j \text{ cov } (r_i , r_j)/(\sum_i s_i^2)^2 = \mathbf{s}'\mathbf{P}\mathbf{s}(\mathbf{s}'\mathbf{s})^{-2} \tag{10}$$

which, for equal selective values, reduces to

$$V(b_I) = \frac{1}{s^2 t^2}\,[t\sigma_d^2 + \sigma_e^2 + h^2\sigma_e^2]. \tag{11}$$

Clearly $V(b_I)$ exceeds $V(b_C)$ (equation 7) if σ_e^2 is of similar magnitude to σ_d^2 . As t increases $V(b_C)$ approaches $1.2\sigma_d^2/s^2t$ and $V(b_I)$ approaches σ_d^2/s^2t, so the latter will be smaller in long term experiments.

iii) *Ratio of total response to total selection differential* (b_R)

The simplest linear estimator, and the one adopted by Hill [1971] for considering the efficiency of alternative designs, is the ratio of the total response to the total selection differential applied over the t generations, namely

$$b_R = R_t/S_t , \tag{12}$$

which has sampling variance

$$V(b_R) = V(R_t)/S_t^2 . \tag{13}$$

When selection differentials are constant, $b_R = b_I$, and thus $V(b_R) = V(b_I)$.

iv) *Maximum likelihood estimator* (b_L)

With divergent selection, two parameters, h^2 and σ^2, have to be estimated. Of these σ^2 can be estimated from the pooled variance between individuals within populations each generation. There is one population in the initial generation and two in each of the remaining t generations, so $(2t + 1)(M - 1)$ D.F. are available for the estimate of σ^2. In most practical situations this should permit a fairly precise estimate to be obtained, so for the ML estimator of h^2, denoted b_L , we shall assume that σ^2 is known without error. The

variance of b_L may thus be biased downwards, but since we shall find that b_L is still little better than the linear estimators in most situations and unsatisfactory in others this bias is clearly not important.

The likelihood (L) may be expressed in terms of the individual generation or cumulative responses and selection differentials. The former has been used since the appropriate variance-covariance matrix, $\mathbf{P}$, is tridiagonal and some reduction in computation is possible. We have

$$\log L = -\tfrac{1}{2}t \log 2\pi - \tfrac{1}{2}\log |\mathbf{P}| - \tfrac{1}{2}(\mathbf{r} - h^2\mathbf{s})'\mathbf{P}^{-1}(\mathbf{r} - h^2\mathbf{s}) \qquad (14)$$

The equation for b_L given by the first derivative of (14) is non-linear and has not been solved explicitly. The large-sample variance of b_L has been obtained from

$$V(b_L) = [-E(d^2 \log L/dh^4)]^{-1}.$$

Noting that $d\mathbf{P}^{-1}/dh^2 = -\mathbf{P}^{-1}(d\mathbf{P}/dh^2)\mathbf{P}^{-1}$, we have

$$V(b_L) = 2\left\{|\mathbf{P}|^{-1}\frac{d^2|\mathbf{P}|}{dh^4} - |\mathbf{P}|^{-2}\left(\frac{d|\mathbf{P}|}{dh^2}\right)^2 + 2\mathbf{s}'\mathbf{P}^{-1}\mathbf{s} + E\left[(\mathbf{r} - h^2\mathbf{s})'\mathbf{P}^{-1}\left(2\frac{d\mathbf{P}}{dh^2}\mathbf{P}^{-1}\frac{d\mathbf{P}}{dh^2} - \frac{d^2\mathbf{P}}{dh^4}\right)\mathbf{P}^{-1}(\mathbf{r} - h^2\mathbf{s})\right]\right\}^{-1}, \qquad (15)$$

where, for any $t \times t$ matrix $\mathbf{A}$ with elements a_{ij},

$$E[(\mathbf{r} - h^2\mathbf{s})'\mathbf{A}(\mathbf{r} - h^2\mathbf{s})] = \sum_i \sum_j a_{ij}p_{ij} .$$

The derivatives of $\mathbf{P}$ and $|\mathbf{P}|$ can be obtained explicitly by differentiating the terms of $\mathbf{P}$ given by (2).

4. COMPARISON OF SAMPLING VARIANCES OF ESTIMATORS

Although it is possible to conduct a selection experiment in which the numbers recorded and selected remain constant every generation, the selection differentials will fluctuate by chance about a mean value of s, dependent on M, N and σ, which can be obtained from tables of order statistics. However, for simplicity most of the comparisons of efficiency are made under the assumption that s takes this constant value. For a specified fraction of the population selected, the selection differentials are proportional to the phenotypic standard deviation, σ. Since σ_s^2 and σ_d^2 are proportional to σ^2 it is clear from equation (8), for example, that the values of the sampling variances, $V(b)$, do not depend differentially on σ. In the following examples a value of $M = 100$ is used, with the highest (or lowest) ranking 5, 10, 20 or 40 selected each generation. However the expected value of s depends primarily on the proportion selected, N/M, rather than on N or M separately. For example, for $p = 0.1$, and $M = 20$, 50, 100 and ∞ then $s = 1.638\sigma$, 1.705σ, 1.730σ and 1.755σ, respectively. Therefore, for a specified proportion selected, the variances of the estimators are very nearly inversely proportional to M, the number recorded each generation.

i) *Equal selection differentials*

The sampling variances of the alternative estimators are compared in Figure 1, in which the proportion selection ($p = N/M$) is held constant and the length of the experiment (t) varied, and in Figure 2 where t is held con-

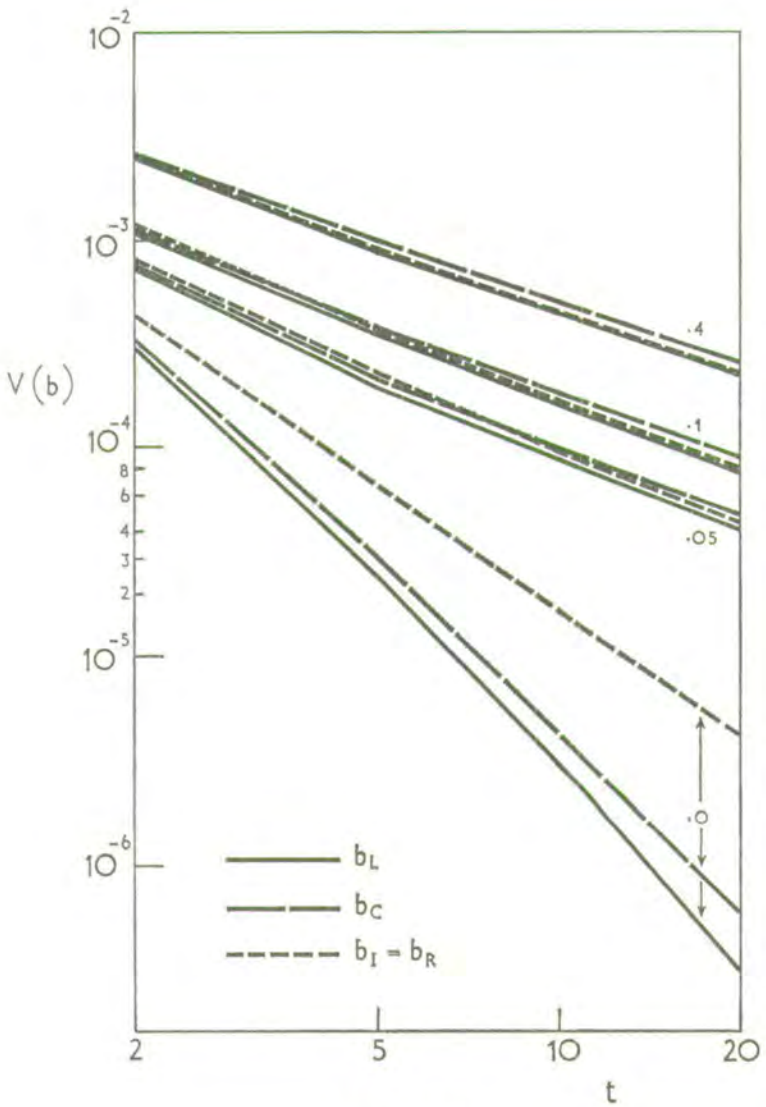

FIGURE 1

SAMPLING VARIANCE OF ALTERNATIVE ESTIMATORS OF REALISED HERITABILITY WITH DIVERGENT SELECTION AND EQUAL SELECTION DIFFERENTIALS FOR $N = 10$, $M = 100$ AND A RANGE OF VALUES OF t AND h^2. (b_L : MAXIMUM LIKELIHOOD, b_C: CUMULATIVE RESPONSES AND SELECTION DIFFERENTIALS, b_I: INDIVIDUAL GENERATION RESPONSES AND SELECTION DIFFERENTIALS, b_R: RATIO OF TOTAL RESPONSE TO TOTAL SELECTION DIFFERENTIAL.

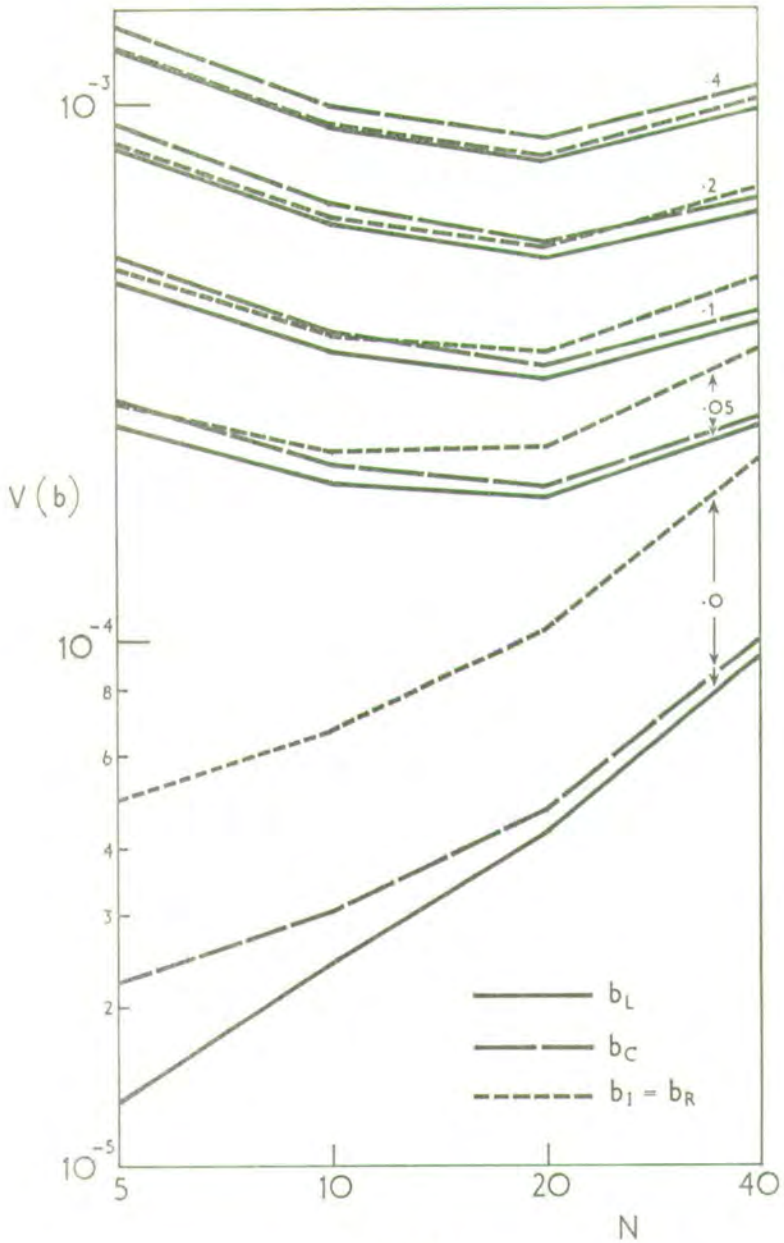

FIGURE 2

As Figure 1, but $t = 5$ for a range of values of N and h^2.

stant and p varied. In each case $M = 100$. It is clear from both graphs that, except where h^2 is very close to zero, the linear estimators b_C and b_I or b_R give almost as good estimates of realised heritability as the ML estimator. For example, with $p = 0.1$, h^2 in the range 0.05 to 0.6 and t from 2 to 20 generations, neither $V(b_C)$ nor $V(b_I)$ exceeds $V(b_L)$ by more than

about 17%. When h^2 is small the regression of cumulative response on cumulative selection differential, b_C , is the best of the linear estimators, and, for $h^2 = 0$, is much poorer than the ML estimator only when p is small and t is large. The estimators b_I or b_R are better than b_C only when t or h^2 is high; even then the differences in sampling variance are small.

The apparent superiority of the ML estimator at very small h^2 values is rather surprising, since if it is known a priori that $h^2 = 0$, then b_C is itself the ML estimator. Further examination reveals, however, that b_L is not useful when h^2 is small. Because of the difficulty of finding an explicit formula for b_L from (14) we illustrate with some simulated sampling experiments from a model described exactly by equations (1), with $M = 100$, $N = 10$ and $t = 5$. With the results simulated with true h^2 values of 0.4 or 0.6, b_L is not noticeably biased. However with $h^2 = 0.1$ and $h^2 = 0.05$, the b_L values (found by trial and error from (14)) in each of 10 replicated experiments at both h^2 levels were lower than the corresponding values of b_C . Since the latter are unbiased, we can infer that those for b_L are biased. Also, with $h^2 = 0.0$, no local minima were found for b_L . Further examination of (14) reveals that there is a discontinuity in the log likelihood at slightly negative values of b (here b denotes some estimate of heritability, which replaces h^2 in equations (2) and (5)). For example, with $t = 5$, $M = 100$, $N = 10$ and $b \leq 0.008$, then $|\mathbf{P}| < 0$ so that log L does not exist. At higher values of t and lower values of N, $|\mathbf{P}|$ is negative for values of b closer to zero, although still negative. The large sample variances of the ML estimator therefore appear to be spurious at low h^2 values, but they have been left in Figures 1 and 2 for completeness. It is clear that maximum likelihood can not be used when the true heritability is low, and we see from the figures that one or more of the linear estimators are efficient at higher values of heritability.

The number of generations and number of animals which can be recorded each generation and are available for a realised heritability estimate may be fixed, so that the experimenter is only able to control the proportion selected each generation. For any value of h^2 there is a value of p at which $V(b)$ is minimised, and this design problem was discussed in the earlier paper, in which b_R was used as the estimator (Hill [1971]). However it is clear from Figure 2 that the minimum values of $V(b)$ for each estimator are found at roughly the same values of p, so the optimal design is scarcely influenced by the estimator adopted, and the results given by Hill [1971] can be used without modification.

ii) *Variable selection differentials*

The effect of variable selection differentials on the relative efficiency of the alternative estimators was investigated for a few sets of parameters to check whether the general pattern shown in Figures 1 and 2 would still hold. Groups of t selection differentials were sampled independently from a normal distribution with mean s and variance c^2s^2, where s is the expected differential when N individuals are selected from M and c is the coefficient of variation of the selection differential. For each group of selection differentials the sampling

variance of each estimator and ratios such as $V(b_C)/V(b_L)$ were computed. The sampling of selection differentials was replicated, and averages of ratios over 50 replications are given in Table 1, together with the appropriate values for equal selection differentials. It is clear that the relative efficiency of the estimators is little changed by the inclusion of variation in selection differentials. Also, the coefficients of variation used in Table 1 are larger than might be expected in practice. For example, with $p = 0.1$, $c = 0.32$ for $M = 10$, $c = 0.22$ for $M = 20$ and is approximately proportional to $1/M^{1/2}$. Although the distribution of s is not normal, the assumption of normality in the simulation should not have introduced any important bias.

When selection differentials are unequal the estimators b_R and b_I are no longer the same, and the ratio of their sampling variances are also given in Table 1, where it is apparent that differences between $V(b_R)$ and $V(b_I)$ are small. Their relative magnitudes can be found under some assumptions. Consider an experiment in which the observed selection differentials have mean $\bar{s}$ and coefficient of variation c. Let $d_i = s_i - \bar{s}$, and, subject to the restraint $\sum_i d_i = 0$, assume the d_i are uncorrelated with each other and the errors of the responses r_i ; then $E(d_i^2) = c^2\bar{s}^2$ and $E(d_i\, d_j) = -c^2\bar{s}^2/(t-1)$.

TABLE 1

RELATIVE EFFICIENCY OF ALTERNATIVE ESTIMATORS OF REALISED HERITABILITY WITH DIVERGENT SELECTION WHEN SELECTION DIFFERENTIALS VARY

The mean selection differential is $\bar{s}$, its coefficient of variation is c and $M = 100$, $\sigma^2 = 1$ throughout. (b_L: maximum likelihood, b_C: cumulative responses and selection differentials, b_I: individual responses and selection differentials, b_R: ratio of total response to total selection differential).

t	$\bar{x}$	$\bar{s}$	h^2	c	$V(b_L)$ x10^4	$\frac{V(b_C)}{V(b_L)}$	$\frac{V(b_I)}{V(b_L)}$	$\frac{V(b_R)}{V(b_L)}$	$\frac{V(b_I)}{V(b_R)}$
5	10	1.730	.05	.0	1.97	1.076	1.150	1.150	1.000
				.4		1.103b	1.310c	1.167b	1.128c
			.4	.0	9.06	1.100	1.006	1.006	1.000
				.4		1.153c	1.016a	1.049b	0.970b
10	5	2.018	.05	.0	1.18	1.170	1.092	1.092	1.000
				.4		1.212b	1.180b	1.125b	1.049b
			.4	.0	6.18	1.144	1.002	1.002	1.000
				.4		1.212c	1.006a	1.059b	0.950b
5	40	0.958	.05	.0	2.51	1.037	1.395	1.395	1.000
				.4		1.039a	1.555c	1.391b	1.121c
			.4	.0	9.79	1.138	1.025	1.025	1.000
				.4		1.145c	1.082b	1.067b	1.015b

Magnitude of standard errors

(a) S.E. ≤ 0.001, (b) $0.001 <$ S.E. ≤ 0.01, (c) $0.01 <$ S.E. ≤ 0.03

From (10)

$$V(b_I) = [(\bar{s}^2 + \sum_i \sum_j d_i\, d_j)\ \mathrm{cov}\ (r_i\ ,\ r_j)]/(t\bar{s}^2 + \sum_i d_i^2)^2$$

$$= [(\bar{s}^2 + \sum_i \sum_j d_i\, d_j)\ \mathrm{cov}\ (r_i\ ,\ r_j) - 2 \sum_i d_i^2/t]/t^2\bar{s}^4$$

approximately, if we assume c is small. Rewriting (13) as

$$V(b_R) = \textstyle\sum_i \sum_j \mathrm{cov}\ (r_i\ ,\ r_j)/t^2\bar{s}^2$$

and substituting, it can be shown that

$$V(b_I) = V(b_R) + \frac{c^2}{t^2\bar{s}^2}\,[(2t - 3)\sigma_e^2 - t\sigma_d^2] \qquad (16)$$

approximately. Therefore b_R is the better estimator if $\sigma_d^2 < (2 - 3/t)\sigma_e^2$, or $p > h^2/(2 - 3/t)$, approximately, but it is clear from (16) that differences in efficiency cannot be large.

iii) *Variable population sizes*

For simplicity we have assumed that the numbers recorded and selected are the same every generation. Modification of equations (1) or (2) to incorporate changes in M or N are straightforward, and some relevant formulae are given by Hill [1971]. For example, if M_i and N_i are the numbers recorded and selected at generation i, the equation for $V(r_i)$ from (2) becomes

$$V(r_i) = 2\sigma^2[h^2(1 - h^2)/N_{i-1} + (1 - h^2)^2/M_{i-1} + 1/M_i] \qquad i = 2, \cdots, t.$$

A few sets of M_i and N_i have been chosen for illustration, and the variances of the alternative estimators are compared in Table 2. Although the simple linear estimators are a little poorer than the ML estimator with some of the chosen parameters, the differences are never large. The estimator b_R seems most robust against changes in population size.

5. ESTIMATION OF SAMPLING VARIANCES

We have noted that commonly employed estimates of sampling variance of the regression of cumulative response on cumulative selection differential are biased. We shall now investigate the magnitude of the bias, and suggest unbiased methods of estimating the variance of this and other linear estimators.

i) *Standard regression methods*

In the usual method of estimating the sampling variance of b_C (Falconer [1960], Richardson *et al.* [1968]) the estimator, denoted $U(b_C)$, is given by

$$U(b_C) = (\sum_i R_i^2 - b_C \sum_i R_i S_i)/(t - 1) \sum_i S_i^2\ . \qquad (17)$$

For this to be an unbiased estimator of $V(b_C)$ the variance-covariance matrix of cumulative responses must be of scalar form, but as we have shown, it

TABLE 2

RELATIVE EFFICIENCY OF ALTERNATIVE ESTIMATORS OF REALISED HERITABILITY WITH DIVERGENT SELECTION WHEN POPULATION SIZES CHANGE.

In each example the sizes and selection differentials are the same for both up and down selection, and $\sigma^2 = 1$, $t = 5$. Changes are made from a basic design of $M_i = 100$, $N_i = 10$, $s_i = 1.730$ for $i = 1, \cdots, 5$ and $M_6 = 100$. (b_L: maximum likelihood, b_C: cumulative responses and selection differentials, b_I: individual responses and selection differentials, b_R: ratio of total response to total selection differential).

	Design	h^2	$\frac{V(b_L)}{x10^4}$	$\frac{V(b_C)}{V(b_L)}$	$\frac{V(b_I)}{V(b_L)}$	$\frac{V(b_R)}{V(b_L)}$
(i)	Basic	.05	1.97	1.076	1.150	1.150
		.4	9.06	1.100	1.006	1.006
(ii)	$M_6 = 1000$	.05	1.60	1.246	1.039	1.039
		.4	8.47	1.162	1.005	1.005
(iii)	$M_2 = 20$, $s_2 = 0.767$	.05	2.55	1.153	1.835	1.127
		.4	11.62	1.261	1.152	1.040
(iv)	$M_2 = 20$, $N_2 = 2$,	.05	3.06	1.419	1.156	1.180
	$s_2 = 1.638$	.4	12.53	1.746	1.264	1.302
(v)	As (iv) and $M_6 = 1000$	.05	3.13	1.782	1.827	1.221
	$M_4 = 20$, $s_4 = 0.767$	.4	15.80	1.771	1.407	1.295

is not. The expected value of $U(b_C)$ may be written

$$E[U(b_C)] = [\sum_i V(R_i) - \sum_i S_i^2 V(b_C)]/(t-1) \sum_i S_i^2 , \qquad (18)$$

where $V(b_C)$ is given by (7). With equal selection differentials (18) reduces to

$$E[U(b_C)] = \frac{6}{s^2 t(t+1)(2t+1)} \left[\frac{t+2}{10} \sigma_d^2 + \sigma_e^2 + \frac{t}{2(2t+1)} h^2 \sigma_e^2 \right] \qquad (19)$$

which can be compared directly with (8). The coefficients of σ_e^2 is the same in $E[U(b_C)]$ as in $V(b_C)$, even when selection differentials are unequal, so there is no bias in the estimate of the sampling variance if $h^2 = 0$. However the coefficient of σ_d^2 is much smaller in the former: in $E[U(b_C)]$ the term in σ_d^2 is of order $1/t^2$, whereas it is of order $1/t$ in $V(b_C)$. The magnitude of the bias is illustrated in Figure 3 for equal selection differentials and the same model as in Figures 1 and 2. It can be seen that the sampling variance of the realised heritability estimate from the regression of cumulative response on cumulative selection differential may be one-tenth or less of the correct value.

On the other hand, a similar analysis of the responses and selection differentials from individual generations over-estimates the sampling variance of the estimator b_I since the negative correlation between successive responses is excluded. If the error structure is ignored, the standard estimate of the

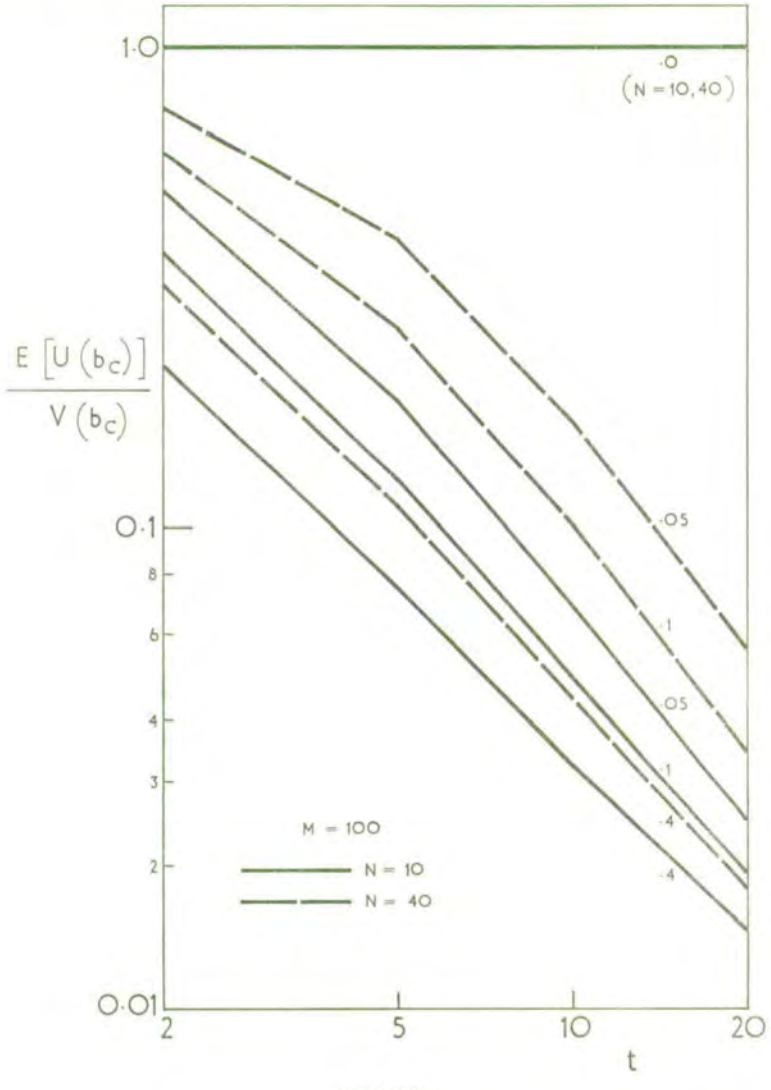

FIGURE 3

COMPARISON OF THE VARIANCE, $V(b_C)$, OF THE REGRESSION OF CUMULATIVE RESPONSE ON CUMULATIVE SELECTION DIFFERENTIAL WITH THE EXPECTED VALUE OF THE ESTIMATE OF VARIANCE OBTAINED BY STANDARD METHODS, $E[U(b_C)]$, FOR DIVERGENT SELECTION WITH EQUAL SELECTION DIFFERENTIALS, $M = 100$ AND SEVERAL VALUES OF h^2.

variance is

$$U(b_I) = (\sum_i r_i^2 - b_I \sum_i r_i s_i)/(t - 1) \sum_i s_i^2$$

when, for equal selection differentials,

$$E[U(b_I)] = \frac{1}{t^2 s^2} [t\sigma_d^2 + (2t + 1)\sigma_e^2 + h^2\sigma_e^2] \qquad (20)$$

The ratio $E[U(b_I)]/V(b_I)$ is given in Figure 4 for the same parameters as in the earlier figures. The bias in the estimate of the sampling variance is serious unless selection is intense and the heritability high, such that $\sigma_d^2 \gg \sigma_e^2$.

ii) *Proposed method*

If a selection experiment has been replicated several times, the sampling variance of the average realised heritability can be estimated from the variance between replicates. This estimate does not require many of the

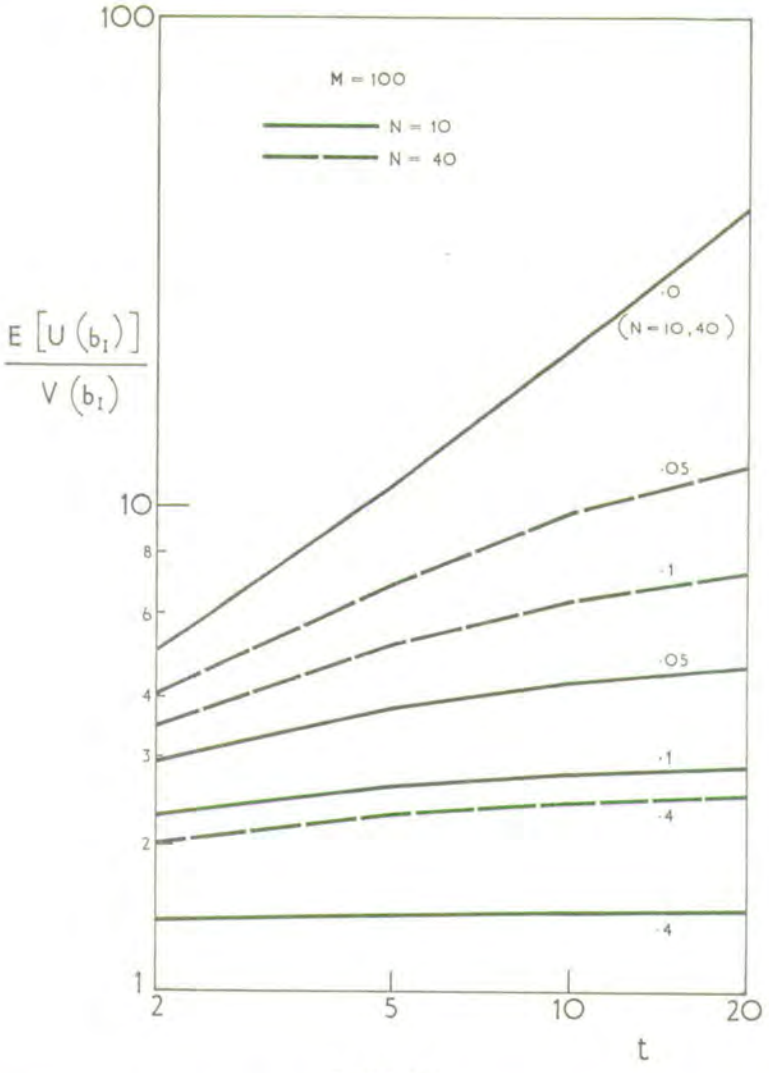

FIGURE 4

COMPARISON OF THE VARIANCE, $V(b_I)$, OF THE REGRESSION OF INDIVIDUAL GENERATION RESPONSES ON SELECTION DIFFERENTIALS WITH THE EXPECTED VALUE OF THE ESTIMATE OF VARIANCE OBTAINED BY STANDARD METHODS, $E[U(b_I)]$, FOR DIVERGENT SELECTION WITH EQUAL SELECTION DIFFERNTIALS, $M = 100$ AND SEVERAL VALUES OF h^2.

assumptions about the model which have been made in our analysis, and should be used wherever possible. However we are more concerned with experiments in which only one, or very few, replicates are available such that the estimate of variance between replicates can not be used. Yet estimates of $V(b)$ are required which are much less biased than those commonly used and discussed above.

The method we propose is as follows: Estimate σ^2 by $\hat{\sigma}^2$ from within populations directly, and estimate h^2 by b_G from (6). The estimates $\hat{\sigma}^2$ and b_G are now used to replace the parameter values σ^2 and h^2 in the appropriate equations (3 and 4) for σ_d^2 and σ_*^2 to obtain estimates of these quantities, and similarly for the variances and covariances of the R_i which are elements of $\mathbf{C}$ (equations (1) and (5)). Then $V(b_G)$ is estimated from (7). This estimate is denoted $\hat{V}(b_G)$. The final calculation is laborious by hand, and a sufficiently precise estimate may be obtained under many circumstances by assuming that the selection differentials take a constant value equal to their mean. Then we replace s by $\bar{s} = S_t/t$ in (8) to obtain a simple estimate of $V(b_G)$.

iii) *Example*

The method is illustrated by an example. The data used were simulated, and sampled from a hypothetical genetic population which behaved exactly as in the model assumed here with $N = 10$, $M = 100$, $\sigma^2 = 100$, $h^2 = 0.4$ and $t = 5$. The results are summarized below:

Generations	(i)	1	2	3	4	5
Cumulative selection differential	(S_i)	31.03	65.09	96.76	132.12	165.22
Cumulative response	(R_i)	14.19	28.21	39.54	57.25	70.12

From an analysis of variance within and between generations, we obtained $\hat{\sigma}^2 = 98.80$ with $11 \times 99 = 1089$ D.F. From (6), $b_G = 0.4257$. Thus from (3), $\hat{\sigma}_d^2 = 5.189$ and from (4), $\hat{\sigma}_*^2 = 1.135$, and, for example, from (1) and (5), $V(R_1) = c_{11} = 5.029$. Substituting into (7) we obtain $\hat{V}(b_G) = 1.12 \times 10^{-3}$.

By contrast, the standard method of estimation of variance (17) gives $U(b_G) = 2.11 \times 10^{-5}$. The correct value of $V(b_G)$, i.e. using the parameter values of σ^2 and h^2, conditional on the set of selection differentials obtained in this experiment, has been found from (7), and is $V(b_G) = 1.10 \times 10^{-3}$. Thus the estimate $\hat{V}(b_G)$ is close to the correct value whilst $U(b_G)$ is much too low.

Assuming that the selection differentials are all constant, with $\bar{s} = 165.22/5 = 33.04$ and substituting in (8), we obtain more simply $\hat{V}(b_G) = 1.11 \times 10^{-3}$, which is very close to the value of $\hat{V}(b_G)$ obtained from (3).

With the parameters of this example it is clear from Figure 1 that the sampling variance of b_I or b_R is lower than that of b_G . For this example, the estimates are $b_I = 0.424$ and $b_R = 0.422$, both very close to b_G . An estimate $\hat{V}(b_I)$ or $V(\hat{b}_R)$ can be obtained in the same way as that described above for $\hat{V}(b_G)$.

iv) *Properties of proposed estimator*

The method gives slightly biased estimates of the sampling variances, since the terms in drift variance are not linear in h^2. We now find the magnitude of this bias and show it to be small.

From (3) and (4), and noting that $\hat{\sigma}^2$ is uncorrelated with any of the linear estimators, b, we have described, it can be shown that

$$E(\hat{\sigma}_d^2) = \sigma_d^2 - 2\sigma^2(1/N - 1/M)V(b) \tag{21}$$

approximately, and

$$E(\hat{\sigma}_e^2) = \sigma_e^2 .$$

Consider, for example, b_R , which has the simplest formula for the sampling variance. From (1) and (13)

$$V(b_R) = [t\sigma_d^2 + \sigma_e^2 + h^2\sigma_s^2]/S_t^2. \tag{22}$$

If the average selection differential is I standard deviations in each direction, then $S_t = 2tI\sigma$, and in most experiments $1 \leq I \leq 2$ (corresponding to $0.4 \geq p \geq 0.05$). Using (21) to evaluate (22) we obtain, as an approximation,

$$E[\hat{V}(b_R)] = V(b_R)\left[1 - \frac{1}{2NtI^2}(1 - p + p/t)\right]$$

and since $2NtI^2 \gg 1$ in any worthwhile experiment, the bias can be ignored.

The estimators of variance, $\hat{V}(b_C)$ or $\hat{V}(b_I)$, apart from being almost unbiased, have, themselves, a much smaller sampling variance than do $U(b_C)$ or $U(b_I)$ from standard methods. The latter are obtained with only the $t - 1$ D.F. among the mean responses, whereas $\hat{V}(b_C)$ and $\hat{V}(b_I)$ utilize σ^2 estimated with $(2t - 1)(M - 1)$ D.F. and the regression coefficient itself. The differences in variance can be shown algebraically. However it is sufficient here to use some simulated sampling experiments for illustration, with $t = 5$, $M = 100$ and $N = 10$, a model described by (1) and 10 replicates at each heritability level. For $h^2 = 0.4$ the range of values of $V(b_C) \times 10^4$ obtained were 9.1 to 11.2, whereas $U(b_C) \times 10^4$ ranged from 0.19 to 3.3, and for $h^2 = 0.1$ the ranges were 3.1 to 4.3 and 0.12 to 0.83 respectively. With equal selection differentials the true values of $V(b_C) \times 10^4$ are 9.97 for $h^2 = 0.4$ and 3.75 for $h^2 = 0.1$ (Figure 1).

6. EXTENSIONS TO THE MODEL

The model we have used is restricted in many ways. The extension beyond divergent selection will be considered in another paper. Here we consider problems of effective population size.

i) *Two sexes*

The results have been given for a model with selection in only one sex merely for simplicity, and the extension to the practical situation of two sexes is straightforward (Hill [1971]). For example, let us assume that with

divergent selection M_m males and M_f females are recorded, and N_m and N_f are selected every generation, that random mating with Poisson family sizes is practiced and that the variances are the same in both sexes. It can then be shown that, for example,

$$\sigma_d^2 = \tfrac{1}{2}\sigma^2[h^2(1 - h^2)(1/N_m + 1/N_f) + h^4(1/M_m + 1/M_f)],$$
$$\sigma_e^2 = \tfrac{1}{2}\sigma^2(1 - h^2)(1/M_m + 1/M_f),$$

and these values can be inserted directly in our formulae. If we define effective sizes in the usual way,

$$\frac{1}{M_e} = \frac{1}{4M_m} + \frac{1}{4M_f}, \qquad \frac{1}{N_e} = \frac{1}{4N_m} + \frac{1}{4N_f},$$

the equations in the earlier sections, including (3) and (4) for σ_d^2 and σ_e^2 can be used directly, with M_e replacing M, N_e replacing N and the mean selection differential in the two sexes used for s_i.

ii) *Non-Poisson family sizes*

We have assumed in the analysis that, even with a monecious model, the actual and the variance effective population numbers are the same. This requires that family sizes should be Poisson distributed or, strictly, multinomial since the total number is fixed (Crow and Kimura [1970]). This assumption will not hold if selection is practiced within families, or family sizes prior to selection are not Poisson in form. Selection itself reduces the effective size since relatives have a correlated performance on the quantitative trait and so, even with mass selection, the variance of family sizes is increased by selection (Robertson [1961]). In each case the effective size should be substituted for the actual size in the formulae given in this paper.

Rather similarly, we have used σ^2/M as the variance of the mean phenotypic value. Essentially, this requires that all families have one individual, whereas if there is a hierarchical mating structure with d dams per sire and n progeny per dam, this variance for an additive trait should be increased to

$$\frac{\sigma^2}{M}[\tfrac{1}{4}ndh^2 + \tfrac{1}{4}nh^2 + 1 - \tfrac{1}{2}h^2];$$

but this modification ignores the effect of selection among the parents, which tends to reduce the variance between families for the selected trait. To some extent these effects cancel, so there is probably little advantage in correcting the variances given in our earlier equations.

7. DISCUSSION

Our objective in this paper has been to find an estimator for realised heritability and its sampling variance when the selection experiment is being used to measure heritability in the base population. We have made no attempt to study whether, for example, the heritability changes during the course of the experiment. The model is based on rather crucial assumptions which

have been discussed more fully previously (Hill [1971]). We assume that the variance within populations (σ^2) and the increment in drift variance (σ_d^2) remain constant, and that genotypic and phenotypic values of progeny of selected parents are normally distributed. These approximations should hold reasonably well if the experiment is of short duration, the final inbreeding level (F) is small and the quantitative trait under selection is not determined by genes of very large effect. Recently Bulmer [1971] has found that some decline in variance is expected since a correlation of gene frequencies is induced by selection. With unlinked loci, the variance can be shown to be reduced by a total of about $h^4\sigma^2/2$ for a wide range of selection intensities, with most of the decline occurring within the first two generations of selection. Thus the model assumptions hold more closely when heritabilities are low.

In the basic model all gene action is assumed to be additive. With dominance the accumulated drift variance (variance between replicates) is also a function of gene frequency. However, from formulae given by Crow and Kimura [1970 p. 343] it can be shown that the coefficient of F in the between-lines variance is $h^2\sigma^2$ for an unselected population, just as for additive genes. Although the term in F does not include the initial gene frequency, higher order terms in F do, but so long as we retain our assumption of a small total inbreeding we can ignore these terms in F^2, F^3 etc. For recessive genes of low frequency, q, the term in F^3 has the lowest order term in q, and could thus be the major term in the drift variance for such genes (Robertson [1952]). But, in a trait affected by a mixture of additive and dominant genes at varying frequencies, these recessives at low frequency, with variance proportional to $q^3(1 - q)$, will contribute only a trivial fraction of the additive variance in the base population, and can be ignored. Thus our model should still be appropriate when there is dominance, and with divergent selection any effects of inbreeding depression are eliminated from the difference in response.

Perhaps the most important finding from this study is one which is essentially negative. Despite the rather involved error structure of the responses, itself dependent on the unknown parameter, h^2, we find that simple linear estimators of realised heritability are almost as efficient as a ML estimator over most of the relevant range of parameters, and, unlike the ML estimator, are unbiased over all this range. We can therefore recommend continued use of the regression of cumulative response on selection differential as the estimator under most circumstances. Unless the heritability is very low, or the experiment of very short duration, most of the variance in later generations is contributed by genetic drift. Then the most simple estimator, the ratio of total response to total selection differential applied during the experiment (b_R), is highly efficient. Standard methods of estimating the variance of realised heritabilities from single experiments are not satisfactory, however, but we have suggested some methods of improvement. With these it can be a straightforward procedure to estimate the variance for each method and use the one with smallest sampling variance.

ACKNOWLEDGMENTS

I am indebted to Miss K. Paver for considerable technical assistance. This research was supported in part by Grant No. GM 13827 from the U. S. National Institutes of Health while the author was at the Statistical Laboratory, Iowa State University, Ames.

ESTIMATION DES HERITABILITES OBTENUES A PARTIR D'EXPERIENCES DE SELECTION. I. SELECTION DIVERGENTE

RESUME

On compare des méthodes d'estimation de l'héritabilité obtenue à partir d'expériences de sélection. Pour des plans d'expérience utilisant la sélection divergente, on donne des formules pour la variance d'échantillonnage de quelques simples estimateurs linéaires de l'héritabilité obtenue, telle la régression de la réponse cumulée sur la sélection différentielle cumulée. Bien que la structure de variance-covariance des réponses dépende de l'héritabilité, on trouve que pour les plus utiles combinaisons de paramètres, ces estimateurs linéaires sont presque aussi efficaces qu'un estimateur du maximum de vraisemblance et peuvent être recommandés pour l'utilisation pratique. On montre que les méthodes standards de calcul de la variance de ces estimateurs sont très biaisées, comparativement à la régression de la réponse cumulée sur la sélection différentielle cumulée. On décrit des méthodes d'estimation de la variance à partir de données expérimentales, méthodes presque sans biais.

REFERENCES

Bulmer, M. G. [1971]. The effect of selection on genetic variability. *Amer. Natur. 105*, 201–11.

Crow, J. F. and Kimura, M. [1970]. *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Dickerson, G. E. [1969]. Techniques for research in quantitative animal genetics. In *Techniques and Procedures in Animal Production Research* Amer. Soc. Anim. Sci., New York, 36–79.

Falconer, D. S. [1954]. Asymmetrical responses in selection experiments. *Un. int. Sci. biol.* No. 15, 16–41.

Falconer, D. S. [1960]. *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh.

Hill, W. G. [1971]. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics 27*, 293–311.

Richardson, R. M., Kojima, K. and Lucas, H. L. [1968]. An analysis of short term selection experiments. *Heredity 23*, 493–506.

Robertson, A. [1952]. The effects of inbreeding on the variation due to recessive genes. *Genetics 37*, 189–207.

Robertson, A. [1961]. Inbreeding in artificial selection programmes. *Genet. Res. 2*, 189–94.

Received July 1971, *Revised November* 1971

Key Words: Genetics; Realised heritability estimation; Design of selection experiments; Regression estimation; Maximum likelihood.

12

Estimation of realised heritabilities from selection experiments

II. Selection in one direction

by

William G. Hill

Reprinted from
BIOMETRICS *28*, 767–780
September 1972

ESTIMATION OF REALISED HERITABILITIES FROM SELECTION EXPERIMENTS. II. SELECTION IN ONE DIRECTION

WILLIAM G. HILL

Institute of Animal Genetics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN Scotland

SUMMARY

The sampling variances of alternative estimators of realised heritability are compared for experiments in which selection is practiced in one direction. The analysis is undertaken for two types of design: where a control is maintained and where it is not, and the criteria for evaluating the utility of a control are discussed. When no control population is kept, the best linear estimator is usually the regression of cumulative response on cumulative selection differential, and this estimator is generally satisfactory even when a control is maintained. Methods of estimating the sampling variance of the realised heritability and the variance due to common environment effects are described and discussed.

1. INTRODUCTION

In previous papers we have investigated both the efficiency of selection experiments for estimating realised heritabilities, and the methods of obtaining these estimates in experiments in which divergent selection is practiced (Hill [1971; 1972a]). More commonly, selection is carried out in a single direction and a control population may be maintained. We consider the analyses of such designs in this paper. If there is a control, common environmental effects are eliminated in the same way as with divergent selection, so little new analysis is required; such theory as is necessary is deferred to section 5. We concentrate first on experiments in which no control population is kept such that environmental effects common to all individuals in the selected line influence the estimate of response and its variance. We shall assume, however, that there is no directional trend in the environment, so that unbiased estimates of realised heritability can be obtained.

Where relevant formulae have been derived in one of the earlier papers (Hill [1971; 1972a]) they will be stated here without proof. The important model assumption which we make is that genetic and phenotypic variances remain constant during the experiment, which must therefore be short term.

2. MODEL

A single selected line is maintained without a control and a total of t generations of selection are practiced for some quantitative trait. For simplicity we assume the population is monecious, and every generation M

individuals are recorded, with mean X_i, $i = 0, \cdots, t$, and from these N are selected, with mean Y_i, $i = 0, \cdots, t-1$. In the usual model of two sexes, with M_m and M_f recorded, and N_m and N_f selected, males and females, we can replace N and M by their effective numbers $N_e = (1/4N_m + 1/4N_f)^{-1}$ and $M_e = (1/4M_m + 1/4M_f)^{-1}$ in the following formulae, and use the mean selection differential in the two sexes (Hill [1972a]). The selection differential at generation i is given by

$$s_i = Y_{i-1} - X_{i-1} \qquad i = 1, \cdots, t,$$

and the subsequent response by

$$r_i = X_i - X_{i-1} \qquad i = 1, \cdots, t.$$

We let the heritability of the trait be h^2, and genotypic and phenotypic values be bivariate normally distributed with variances $h^2\sigma^2$ and σ^2 respectively, and correlation h, remaining constant throughout the experiment. Common environmental effects, which influence all individuals in a single generation in the same way, have zero mean, variance σ_c^2 and are uncorrelated, and for the purposes of maximum likelihood (ML) estimation only, are assumed to be normally distributed.

In the base generation we let

$$E(X_0) = \mu$$

and, conditional on the set of selection differentials $s_1, \cdots, s_t$ obtained, implicit in subsequent formulae,

$$\begin{aligned} E(X_i) &= \mu + h^2 \sum_{k=1}^{i} s_k \\ &= \mu + h^2 S_i \end{aligned}$$

where S_i, $i = 1, \cdots, t$, is the cumulative selection differential to generation i and we let $S_0 = 0$.

For this model the genetic drift variance is given by

$$\sigma_d^2 = \sigma^2[h^2(1 - h^2)/N + h^4/M], \tag{1}$$

and we let

$$\sigma_e^2 = \sigma^2(1 - h^2)/M + \sigma_c^2 . \tag{2}$$

By simplification of formulae given by Hill [1971], but ignoring a trivial term of $\frac{1}{2}h^4\sigma^2/M$ which could be included in σ_e^2, we have

$$\begin{aligned} V(X_i) &= i\sigma_d^2 + \sigma_e^2 + h^2\sigma^2/M \qquad i = 0, \cdots, t \\ \operatorname{cov}(X_i, X_j) &= \operatorname{cov}(X_j, X_i) = i\sigma_d^2 + h^2\sigma^2/M \qquad 0 \leq i < j \leq t. \end{aligned} \tag{3}$$

Since the term $h^2\sigma^2/M$ is present in all variances and covariances it is eliminated in formulae based on differences between the X_i. We also have

$$\begin{aligned} V(r_i) &= \sigma_d^2 + 2\sigma_e^2 \qquad i = 1, \cdots, t \\ \operatorname{cov}(r_i, r_{i+1}) &= \operatorname{cov}(r_{i+1}, r_i) = -\sigma_e^2 \qquad i = 1, \cdots, t-1 \\ \operatorname{cov}(r_i, r_j) &= 0 \qquad \text{otherwise.} \end{aligned} \tag{4}$$

Throughout we shall assume that M and N remain constant. However if these have values M_i and N_i at generation i,

$$V(X_i) = \sigma^2 \sum_{k=0}^{i-1} [h^2(1 - h^2)/N_k + h^4/M_k] + \sigma^2/M_i + \sigma_e^2 \qquad i = 0, \cdots, t$$

$$\begin{aligned}\text{cov}\,(X_i, X_j) &= \text{cov}\,(X_j, X_i) \\ &= \sigma^2 \sum_{k=0}^{i-1} [h^2(1 - h^2)/N_k + h^4/M_k] + h^2\sigma^2/M_i \qquad 0 \le i < j \le t\end{aligned}$$

where, if $i = 0$, only the last terms are included.

3. ESTIMATION OF REALISED HERITABILITY

In contrast to the divergent selection case, where there are only two parameters, h^2 and σ^2 to be estimated, we now have an additional two: μ and σ_e^2. However we concentrate on estimation of h^2, and consider the others if necessary for finding the variance of the realised heritability estimate. The four estimators considered by Hill [1972a], namely three linear and the maximum likelihood, all have desirable properties in some situations, and we consider essentially the same ones here.

The ML *estimator* (b_L) of realised heritability cannot be obtained explicitly, and its large sample variance has been found by extension of the method described by Hill [1972a] to include the extra parameters σ_e^2 and μ. We have again assumed that σ^2 is known without error, for it can be estimated within generations with many degrees of freedom. When h^2 is close to zero, ML estimation cannot be used since the likelihood is discontinuous (Hill [1972a]). Thus we have only considered it for h^2 values of at least 0.05, and then only to indicate the relative efficiency of the simple linear estimators.

The linear estimators analyzed are all unbiased. These are as follows:
(i) The *regression of cumulative response on cumulative selection differential* (b_C) was proposed by Falconer [1954] and is

$$b_C = \sum_{i=0}^{t} (S_i - \bar{S})(X_i - \bar{X}) \Big/ \sum_{i=0}^{t} (S_i - \bar{S})^2 \tag{5}$$

where $\bar{S}$ and $\bar{X}$ are the means of the S_i and X_i, over $i = 0, \cdots, t$. In contrast to the case of divergent selection, where the initial mean is zero, and thus known without error, the regression line is not forced through X_0 but passed through $(\bar{S}, \bar{X})$ in the usual way. When $h^2 = 0$, and thus $\sigma_d^2 = 0$, b_C is the least squares estimator of realised heritability. The variance of the estimator is given by

$$V(b_C) = \sum_{i=0}^{t} \sum_{j=0}^{t} (S_i - \bar{S})(S_j - \bar{S})\ \text{cov}\,(X_i, X_j) \Big/ \left[\sum_{i=0}^{t} (S_i - \bar{S})^2\right]^2 \tag{6}$$

where cov (X_i, X_j) is given by (3). With N and M constant,

$$V(b_C) = (AB\sigma_d^2 + \sigma_e^2)A \tag{7}$$

where, for brevity,

$$A = 1 \Big/ \sum_{i=0}^{t} (S_i - \bar{S})^2, \tag{8}$$

$$B = \sum_{i=0}^{t} \sum_{j=0}^{t} (S_i - \bar{S})(S_j - \bar{S}) \min(i, j), \tag{9}$$

and min (i, j) denotes the lower of i or j, e.g. min (3, 2) = 2. Some insight into (7) can be obtained if we assume that the selection differentials are equal each generation, such that $s_i = s$, $S_i = is$. Then (7) reduces to

$$V(b_C) = \frac{12}{s^2 t(t+1)(t+2)} \left[\frac{t^2 + 2t + 2}{10} \sigd^2 + \sigma_e^2 \right] \tag{10}$$

(ii) The *regression of individual generation responses on individual selection differentials* (b_I) is defined by

$$b_I = \sum_{i=1}^{t} s_i r_i \Big/ \sum_{i=1}^{t} s_i^2 ,$$

and is the least squares estimator if $h^2 = 1$ and $\sigma_e^2 = 0$, such that $\sigma_*^2 = 0$. With equal selection differentials, its variance is

$$V(b_I) = \frac{1}{t^2 s^2} [t\sigd^2 + 2\sigma_e^2].$$

(iii) The *ratio of total response to total selection differential* (b_R) is the simplest estimator, and is given by

$$b_R = (X_t - X_0)/S_t .$$

With equal selection differentials, $b_R = b_I$ and thus $V(b_R) = V(b_I)$. When selection differentials vary from generation to generation it can be shown that $V(b_R) < V(b_I)$ if $\sigd^2 < (2 - 3/t)\sigma_e^2$, approximately, as for divergent selection (Hill [1972a]).

The sampling variance of the ML and these linear estimators of realised heritability are compared in Figure 1, in which selection differentials are assumed to be equal each generation. While this cannot be achieved in practice, it can be demonstrated that the relative efficiencies of the estimators are little affected by introducing variability in selection differentials of the magnitude expected in practice. This was illustrated for divergent selection in the previous paper (Hill [1972a]). In Figure 1 the abscissa is taken as σ_e^2 since this is the main parameter included here but excluded in divergent selection. It is clear from the figure that only if h^2 is high, say 0.4 or more, and σ_e^2 close to zero that b_C is less efficient than b_I or b_R , and even under these conditions the differences in efficiency are small. The ML estimator is also little more efficient than b_C , especially when σ_e^2 is large. We have found that these general conclusions are not affected if values of N/M and t, other than those used for Figure 1, are taken. Thus for most practical situations the regression of cumulative response on cumulative selection differential, b_C , should be used, although it is usually little better than the simple estimator, b_R .

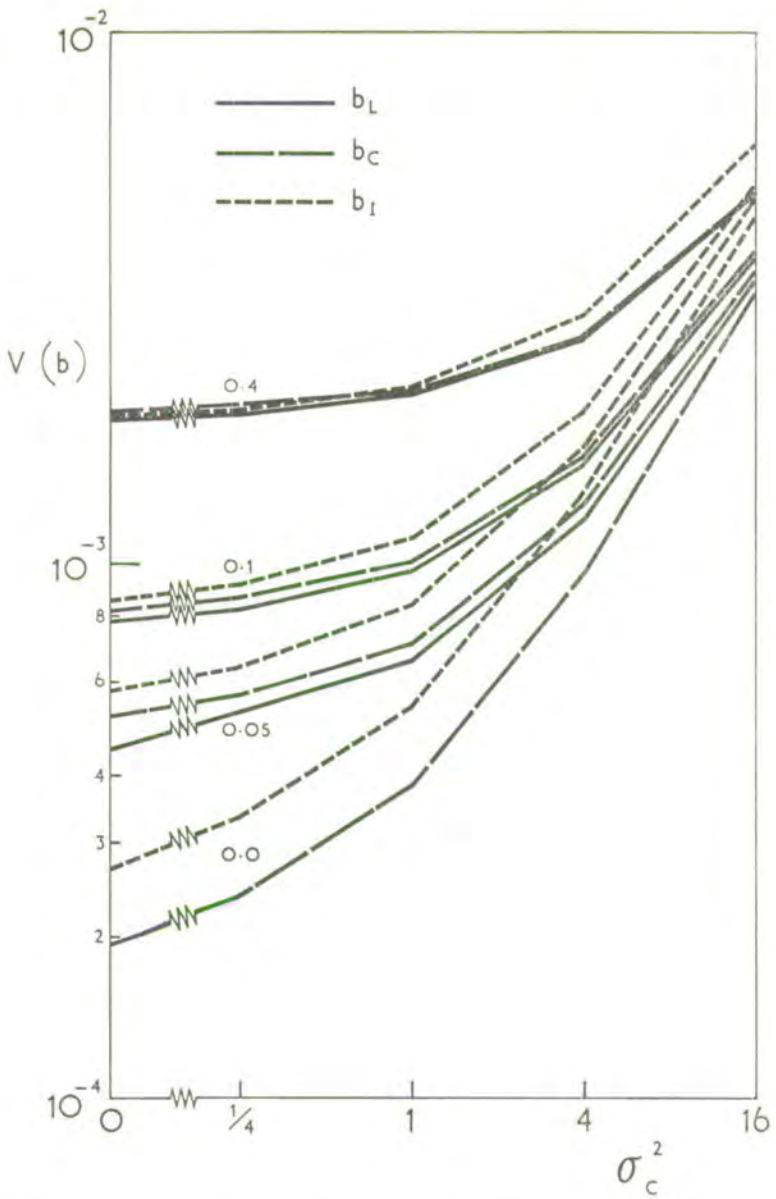

FIGURE 1

SAMPLING VARIANCE OF ALTERNATIVE ESTIMATORS OF REALISED HERITABILITY WITH SELECTION IN ONE DIRECTION, NO CONTROL POPULATION, AND $N = 10$, $M = 100$, $t = 5$ AND A RANGE OF VALUES OF σ_c^2 AND h^2. (b_L: MAXIMUM LIKELIHOOD, b_C: CUMULATIVE RESPONSE AND SELECTION DIFFERENTIALS, b_I: INDIVIDUAL GENERATION RESPONSES AND SELECTION DIFFERENTIALS)

4. ESTIMATION OF SAMPLING VARIANCE

We now consider only b_C, and attempt to provide estimators of its sampling variance from the data available in an experiment. The standard method of estimation (Falconer [1954; 1960], Richardson *et al.* [1968]) is that from linear regression theory in which the X_i are implicitly assumed to be uncorrelated with equal variance, and the relevant error structure is effectively ignored. The estimate of $V(b_C)$ obtained in this way we shall denote by $U(b_C)$, and it can be written

$$U(b_C) = h\left[\sum_{i=1}^{t}(X_i - \bar{X})^2 - b_C \sum_{i=0}^{t}(S_i - \bar{S})(X_i - \bar{X})\right] \Big/ \left[(t-1)\sum_{i=0}^{t}(S_i - \bar{S})^2\right]. \tag{11}$$

The expected value of $U(b_C)$ can be shown to be

$$E[U(b_C)] = \{[t(t+2)/6 - AB]\sigma_d^2/(t-1) + \sigma_e^2\}A, \tag{12}$$

where A and B are defined by (8) and (9). With equal selection differentials, (12) reduces to

$$E[U(b_C)] = \frac{12}{s^2 t(t+1)(t+2)}\left[\frac{t+3}{15}\sigma_d^2 + \sigma_e^2\right]. \tag{13}$$

If $h^2 = 0$ and thus $\sigma_d^2 = 0$, it can be seen from (7) and (12) or (10) and (13) that $E[U(b_C)] = V(b_C)$, and, as we anticipate, the standard estimator is then unbiased. Otherwise it is biased downwards, and the magnitude of bias is illustrated in Figure 2 for the case of equal selection differentials. We see that, particularly when σ_e^2 is small and the experiment is of relatively long duration, $U(b_C)$ can have an expected value 10% or less of $V(b_C)$ and is clearly unsatisfactory as an estimator. With the model of divergent selection σ_e^2 is, in effect, zero and rather larger biases are found than with one way selection (Hill [1972a]).

Estimators of $V(b_C)$ which are almost unbiased can easily be given, however, and we propose a straightforward method here. By an analysis of variance within generations σ^2 can be estimated as, say, $\hat{\sigma}^2$; and using $\hat{\sigma}^2$ and the realised heritability, b_C , the drift variance can also be estimated, as say, $\hat{\sigma}_d^2$, from (1). With divergent selection b_C , $\hat{\sigma}^2$ and $\hat{\sigma}_d^2$ are sufficient to estimate $V(b_C)$, but in this model there remains the unknown σ_e^2 which has to be found from some analysis of variance between generations. We utilize here the analysis giving $U(b_C)$ and note that the coefficient of σ_e^2 (and thus of σ^2 and σ_e^2) in the expected value of $U(b_C)$ is the same as in $V(b_C)$. From (7) and (12) we have

$$V(b_C) = E[U(b_C)] + [AB - (t+2)/6]tA\sigma_d^2/(t-1),$$

and so we use as an estimator

$$\hat{V}(b_C) = U(b_C) + [AB - (t+2)/6]tA\hat{\sigma}_d^2/(t-1). \tag{14}$$

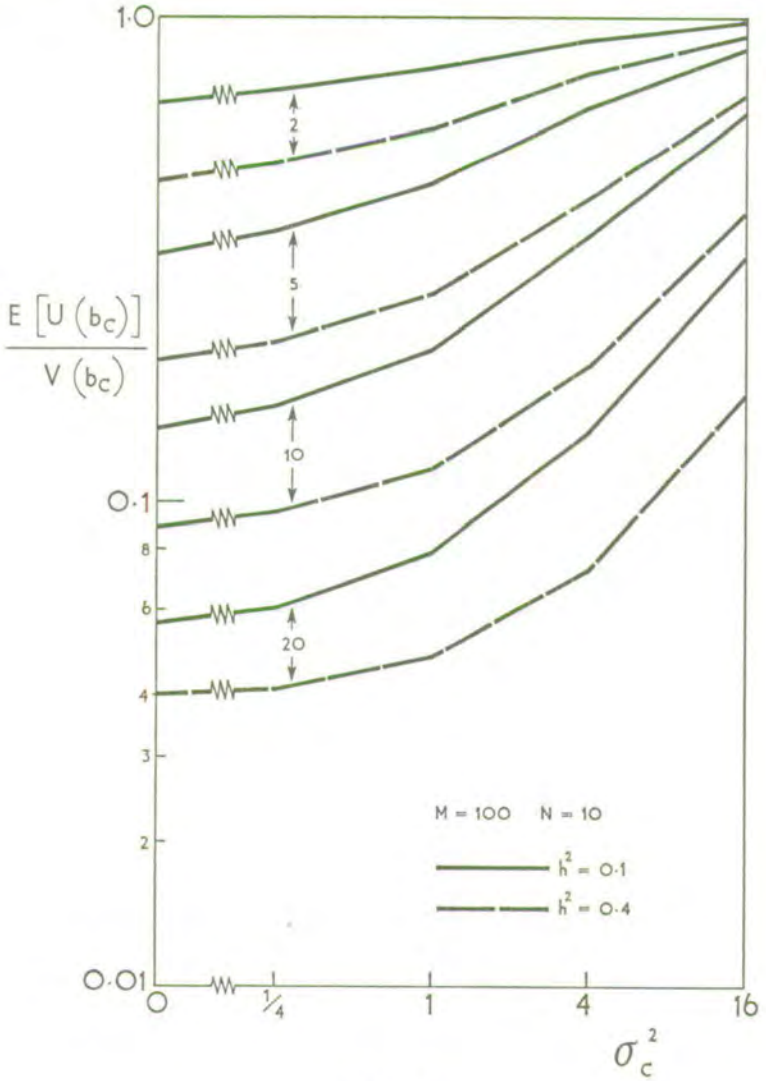

FIGURE 2

COMPARISON OF THE VARIANCE, $V(b_C)$, OF THE REGRESSION OF CUMULATIVE RESPONSE ON CUMULATIVE SELECTION DIFFERENTIAL WITH THE EXPECTED VALUE OF THE VARIANCE OBTAINED BY STANDARD METHODS, $E[U(b_C)]$, FOR SELECTION IN ONE DIRECTION, NO CONTROL POPULATION, EQUAL SELECTION DIFFERENTIALS $N = 10$, $M = 100$, AND A RANGE OF VALUES OF h^2.

Thus to find $\hat{V}(b_C)$ we do not have to estimate σ_e^2 explicitly. Any bias in $\hat{V}(b_C)$ will come from $\hat{\sigma}_d^2$, and we can develop the arguments given by Hill [1972a] to show that this bias must be small. There will be a large sampling variance attached to $\hat{V}(b_C)$ if the experiment is short term so that there are few degrees of freedom between generations. We summarize the method with a numerical example below.

Method and example

To check the methods, data were simulated from the model exactly described by (3), with $t = 5$, $M = 100$, $N = 10$, $\mu = 10$, $\sigma^2 = 100$, $\sigma_e^2 = 4$ and $h^2 = 0.4$. The data after rounding, with deviations of the S_i from their mean, were as follows:

Generation (i)	0	1	2	3	4	5
Cumulative selection differential (S_i)	0.00	20.22	37.26	53.01	72.91	90.75
$S_i - \bar{S}$	−45.69	−25.47	−8.43	7.32	27.22	45.06
Performance (X_i)	9.04	14.19	22.45	20.92	31.71	39.89

(i) Compute

$$A = 1 \Big/ \sum_{i=0}^{t} (S_i - \bar{S})^2 = 1/5631.9 = 1.7756 \times 10^{-4},$$

$$\sum_{i=0}^{t} (S_i - \bar{S})(X_i - \bar{X}) = 2645.6, \qquad \sum_{i=0}^{t} (X_i - \bar{X})^2 = 1256.6.$$

(ii) Estimate h^2 by b_C from (5): $b_C = 0.4698$.

(iii) Compute $U(b_C)$ from (11): $U(b_C) = 6.127 \times 10^{-4}$.

(iv) Estimate σ^2 by $\hat{\sigma}^2$ from an analysis of variance of individual measurements within generations with $(t + 1)(M - 1)$ D.F.: $\hat{\sigma}^2 = 110.9$ with 594 D.F. Where information is available on two sexes, the pooled degrees of freedom should be used.

(v) Estimate σ_d^2 by $\hat{\sigma}_d^2$ from a modification of equation (1), namely

$$\hat{\sigma}_d^2 = \hat{\sigma}^2[b_C(1 - b_C)/N + b_C^2/M],$$

giving $\hat{\sigma}_d^2 = 3.008$ in our example.

(vi) Compute B from (9):

$$\begin{aligned} B &= 1 \times (-25.47)^2 + 2 \times (-8.43)^2 + \cdots + 5 \times (45.06)^2 \\ &\quad + 2[1 \times -25.47 \times (-8.43 + 7.32 + 27.22 + 45.06) + \cdots \\ &\quad + 4 \times 27.22 \times 45.06] \\ &= 20740. \end{aligned}$$

(vii) Compute $\hat{V}(b_C)$ from (14): $\hat{V}(b_C) = 22.92 \times 10^{-4}$.

Our estimate of heritability is $\hat{h}^2 = 0.470 \pm 0.048$. The parameter value of h^2 is 0.4, and the standard error, from Figure 1, would be 0.052 if the selection differentials were all equal to their expected value.

Short-cut method

The computation of B is the only lengthy part of the calculation. So long as the selection differentials do not vary widely from generation to generation the calculations can be considerably reduced by assuming they take a constant value $\bar{s} = S_t/t$, such that $S_i = i\bar{s}$. Then (14) reduces to

$$\hat{V}(b_C) = U(b_C) + \frac{2(3t+4)}{5\bar{s}^2(t+1)(t+2)}\hat{\sigma}_d^2 . \tag{15}$$

In our example, $\bar{s} = 18.15$ and inserting $U(b_C) = 6.127 \times 10^{-4}$ and $\hat{\sigma}_d^2 = 3.008$ we obtain from (15), $\hat{V}(b_C) = 22.65 \times 10^{-4}$. This differs by only 1.2% from the more precise estimate of the previous section, yet the selection differentials ranged from 15.75 to 20.22. Simulations of further replicates of this model, and others with different values of t, h^2 or σ_c^2 have been performed, and the values of $\hat{\sigma}_d^2$ obtained from the short-cut and detailed procedures never deviate by an important amount.

Estimation of σ_c^2

Although estimation of the variance of common environmental effects, σ_c^2, is not carried out in the above analysis, it can be done with little additional computation. Again, equating expectations with observations, we have from (12)

$$\hat{\sigma}_e^2 = U(b_C)/A - [t(t+2)/6 - AB]\hat{\sigma}_d^2/(t-1), \tag{16}$$

and from (2)

$$\hat{\sigma}_c^2 = \hat{\sigma}_e^2 - \hat{\sigma}^2(1 - b_C)/M. \tag{17}$$

Substituting the necessary quantities from our example, we obtain $\hat{\sigma}_e^2 = 1.833$ and $\hat{\sigma}_c^2 = 1.245$. This is considerably below the parameter value of $\sigma_c^2 = 4$ but, of course, few degrees of freedom were available for its estimation.

For the short-cut procedure, we replace (16) by

$$\hat{\sigma}_e^2 = U(b_C)/A - (t+3)\hat{\sigma}_d^2/15$$

which is based on (13). In our example, this gives $\hat{\sigma}_e^2 = 1.847$ and from (17) $\hat{\sigma}_c^2 = 1.257$, in good agreement with the more exact formula.

5. MAINTENANCE OF A CONTROL POPULATION

When a control population is maintained alongside that undergoing selection common environmental effects can be eliminated if there is no interaction between the environment and the control and selected lines' performance. If such an interaction is thought to be present and important the variance which it contributes to the difference in means between selected and control populations takes the place of the common environmental variance in the model of the previous sections and the same analysis can be used unless the interaction variance increases as the difference between the populations increases. In the analysis which follows, we shall assume that there is no genotype-environment interaction.

Several types of control population are possible (Dickerson [1969], Hill [1972b]). We shall only consider non-inbred populations in which breeding individuals are chosen at random and are subject to drift variance. Then two alternatives can be distinguished:

(i) where the control and the selected line are taken from the same base population at generation 0, so the initial mean is known without error; and (ii) where they have a different base and there is error of estimation of the initial mean. Although some genotype-environment interaction is more likely in the latter case we again assume there is none and that the genetic parameters are the same in each population. Thus the two models differ only at generation 0.

Let us assume that in the control an effective number K individuals are recorded (i.e. $1/K = 1/4K_m + 1/4K_f$, where K_m and K_f males and females are recorded) and that the effective size of the breeding population is L. The means, X_t, are now the differences between selected and control performance, and we have

$$\begin{aligned} \sigma_e^2 &= (1 - h^2)\sigma^2/M + \sigma^2/K \\ \sigma_d^2 &= h^2(1 - h^2)\sigma^2/N + h^2\sigma^2/L, \end{aligned} \tag{18}$$

where the error structure is seen to be rather different for the control in which no selection is practiced. The selection differentials are measured only in the selected population.

In case (ii) (control and selected line from different bases) the error structure of equations (3) still holds. Thus the relative efficiency of the estimators is given, in principle, by Figure 1 with $\sigma_e^2 = 0$. The model does not depart far from that of divergent selection, so the more detailed Figures 1 and 2 of Hill [1972a] can also be utilised to indicate differences between, but not absolute values of, the sampling variances of the alternative estimators. Although b_C is not necessarily the best estimator, it is never much poorer than the other linear estimators or a ML estimator. The procedure for estimating $V(b_C)$ should be modified from that given in section 4, since σ_e^2 is eliminated, and the method outlined for divergent selection (Hill [1972a]) can be used. In summary, we can: estimate h^2 by b_C and σ^2 from an analysis within generations, use these to estimate σ_d^2 and σ_e^2 from (18) and thus $V(b_C)$ from (7) or, more simply, (10). The sampling variances of b_I or b_R can be estimated in a similar way.

In case (i) (control and selected lines from the same base), $X_0 = 0$ and $V(X_0) = 0$, so the regression of cumulative response on selection differential can be forced through this point. We denote the estimator b_C^*, given by

$$b_C^* = \sum_{i=1}^{t} X_i S_i \Big/ \sum_{i=1}^{t} S_i^2$$

(in Hill [1972a] this is denoted b_C, where there is no ambiguity since the alternative estimator (5) is not included). Then, in general,

$$\begin{aligned} V(b_C^*) &= \left[\sum_{i=1}^{t} \sum_{j=1}^{t} S_i S_j \min(i, j)\sigma_d^2 \Big/ \sum_{i=1}^{t} S_i^2 + \sigma_e^2 \right] \Big/ \sum_{i=1}^{t} S_i^2 \\ &= \frac{6}{s^2 t(t+1)(2t+1)} \left(\frac{2t^2 + 2t + 1}{5} \sigma_d^2 + \sigma_e^2 \right) \end{aligned} \tag{19}$$

if selection differentials are equal. $V(b_C^*)$ can be found in the same manner as $V(b_C)$ in case (ii) above. The relative efficiency of the alternative estimators b_C^* , b_I or b_R can be obtained from Figures 1 and 2 of Hill [1972a] and the latter two can be up to 20% more efficient than b_C^* when σ_d^2 is larger than σ_e^2 and when t is high. Otherwise b_C^* has least variance.

It is, of course, possible to force the cumulative regression through X_0 (i.e. using b_C^*) even when X_0 is not known exactly, or not pass it through X_0 (i.e. b_C) when it is. By examining the appropriate formulae with equal selection differentials it can be shown that where X_0 is not known without error, b_C should be used since the coefficients of both σ_e^2 and σ_d^2 are at least as high in b_C^* as in b_C . However the result is less clear when X_0 is known exactly, since forcing the regression through this point reduces the contribution to error from σ_e^2 , but increases that from σ_d^2 because more weight is given to the later generations, which have undergone most drift. However, with both $t = 1$ and $t \to \infty$ (with $\sigma_d^2 > 0$) they have the same efficiency, so differences are only seen in intermediate generations; for example, when $t = 5$, $V(b_C^*)/V(b_C) = (\sigma_e^2 + 12.2\sigma_d^2)/(2.02\sigma_e^2 + 11.6\sigma_d^2)$. Then, unless $\sigma_d^2 \gg \sigma_e^2$, it is seen that $V(b_C^*)$ is slightly more efficient, and should probably be used where appropriate, as in the divergent selection case discussed in the previous paper. Similar comparisons are made by Dickerson [1969] of the efficiency of control populations at estimating change per generation in selected lines from common or different base populations.

6. DISCUSSION

In this and the previous papers we have commented on the relative efficiency of divergent selection and schemes of selection in one direction for estimation of realised heritability. We can now summarise the arguments. The sampling variance of an estimate, such as b_C , is seen from our formulae to be proportional to $(\sigma_e^2 + k\sigma_d^2)/s^2$ for equal selection differentials, where k is a constant; for example $k = (t^2 + 2t + 2)/10$ for $V(b_C)$ in equation (10). Some modification is required if we are able to force the regression through some known initial value, but we can show from (11) and (19) that this does not reduce the sampling variance greatly unless the experiment is very short term. Let us assume that a total of M individuals can be recorded each generation. Then with selection in one direction σ_e^2 and σ_d^2 are proportional to $1/M$ if there is no common environmental variance ($\sigma_c^2 = 0$), whereas with divergent selection in two populations of size $M/2$, the variances, σ_e^2 and σ_d^2 , of the differences between the means are proportional to $4/M$. But the value of s^2 is also increased by a factor of 4 with divergent selection, so the two schemes have approximately the same efficiency. However, if $\sigma_c^2 > 0$, selection in one direction without a control immediately becomes less efficient; while if a control population is maintained and resources have to be devoted to it, such that the size of the selected line is less than M, it is still less efficient than divergent selection.

If just one direction of selection is of interest for biological or economic

reasons, the only issue then is whether or not a control population should be maintained. Unfortunately, it is difficult to give any hard and fast rules, but we can at least outline how a decision might be reached. The relevant formulae for the drift and error variances are given in equations (1), (2) and (18), and the relative weights which should be attached to them is k (see above). We consider an example. Let us assume that $t = 5$ and the control comes from a different base, so $k = 3.7$ from (10), and that 10% of the population are selected. Thus with no control population, from (1) and (2)

$$\begin{aligned}\sigma_e^2 + k\sigma_d^2 &= (1 + 36h^2 - 33.3h^4)\sigma^2/M + \sigma_e^2 \\ &= 4.3\sigma^2/M + \sigma_e^2, \quad \text{if} \quad h^2 = 0.1 \\ &= 10.1\sigma^2/M + \sigma_e^2, \quad \text{if} \quad h^2 = 0.4\end{aligned}$$

where a total of M individuals are recorded. If a control is used, let us assume that there are $2M/3$ individuals recorded in the selected line and $M/3$ in the control, and that $M/15$ parents are used in each, with restricted family size in the control, such that its effective size is $2M/15$ (this arrangement is more efficient than partitioning $M/2$ individuals to each population). In this case, from (18)

$$\begin{aligned}\sigma_e^2 + k\sigma_d^2 &= (4.5 + 83.25h^2 - 49.95h^4)\sigma^2/M \\ &= 12.3\sigma^2/M, \quad \text{if} \quad h^2 = 0.1 \\ &= 29.8\sigma^2/M, \quad \text{if} \quad h^2 = 0.4.\end{aligned}$$

Therefore, for it to be more efficient to maintain a control, σ_e^2 must exceed $8\sigma^2/M$ if $h^2 = 0.1$ and up to $20\sigma^2/M$ if $h^2 = 0.4$. Information on the relative sizes of σ^2 and σ_e^2 may be available in laboratories where experiments have previously been undertaken, and the magnitude of σ_e^2 will, of course, depend on the uniformity of the environment. Also, we see that the greater the number of individuals available, the more does the relative efficiency of using a control increase.

We have assumed in the analysis of unidirectional selection experiments that there is no real change in the common environment, such as might be caused by a change in diet or personnel. Any change could bias the heritability estimate if no control were maintained. For example, an environmental trend of z per generation, would, if selected differentials were equal, give a heritability estimate with expected value z/s, and this bias would be important if large relative to the standard error of the estimate. Since such real environmental changes are hard to predict a priori most experimentalists now maintain a control as a reassurance. It is apparent from published data that in many cases they have obtained no benefit in improved precision from having done so, since the test environment has been sufficiently uniform (Hill [1972b]).

In this series of papers we have attempted to construct a theoretical

framework on which to base the design and analysis of selection experiments, and have paid particular attention to methods of estimation of realised heritabilities and their sampling variances. We have not considered all possible permutations of selection schemes, having excluded, for example, selection within families. But such programmes do not appear to introduce any particular analytical problems, since we can specify the appropriate values of the variances σ_e^2 and σ_d^2 and from there the analysis proceeds in the usual way. Nor have we considered estimation of realised genetic correlations (Falconer [1960]) in detail, but it is clear from some preliminary discussion that efficient methods of estimation of realised heritability will be efficient for realised genetic correlations also (Hill [1971]). The analysis has been carried out without reference to any particular set of experimental data, and there is clearly now a need to test our models in practice. The models are subject to many restrictions (Hill [1971; 1972a]) and such analyses should highlight any important deficiencies. In particular, the investigation of data from replicated experiments should provide a worthwhile check.

ACKNOWLEDGMENTS

I am indebted to Miss K. Paver for considerable technical assistance. This research was supported in part by Grant No. GM 13827 from the U. S. National Institutes of Health while the author was at the Statistical Laboratory, Iowa State University, Ames.

ESTIMATION DES HERITABILITES OBTENUES A PARTIR D'EXPERIENCE DE SELECTION. II. SELECTION UNIDIRECTIONNELLE.

RESUME

On compare les variances d'échantillonnage de divers estimateurs de l'héritabilité obtenue, dans le cas d'expériences où l'on pratique la sélection vers une seule direction. L'analyse est effectuée pour deux sortes de plans: ceux qui comprennent un lot témoin et ceux qui n'en comprennent pas, et on discute les critères permettant d'évaluer l'utilité d'un lot témoin. En l'absence de population témoin, le meilleur estimateur linéaire est habituellement la régression de la réponse cumulée sur la sélection différentielle cumulée, et cet estimateur est généralement satisfaisant, même en présence d'un lot témoin. On décrit et on discute des méthodes d'estimation de la variance d'échantillonnage de l'héritabilité obtenue et de la variance dûe aux effets d'un environnement commun.

REFERENCES

Dickerson, G. E. [1969]. Techniques for research in quantitative animal genetics. In *Techniques and Procedures in Animal Production Research.* Amer. Soc. Anim. Sci., New York, 36–79.

Falconer, D. S. [1954]. Asymmetrical responses in selection experiments. *Un. int. Sci. biol.* No. 15, 16–41.

Falconer, D. S. [1960]. *Introduction to Quantitative Genetics.* Oliver and Boyd, Edinburgh.

Hill, W. G. [1971]. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics 27*, 293–311.

Hill, W. G. [1972a]. Estimation of realised heritabilities from selection experiments. I. Divergent selection. *Biometrics 29*, 747–765.

Hill, W. G. [1972b]. Control populations. *Anim. Breed. Abstr. 40*, 1–15.

Richardson, R. M., Kojima, K. and Lucas, H. L. [1968]. An analysis of short term selection experiments. *Heredity 23*, 493–506.

Received July 1971, *Revised November* 1971

Key Words: Genetics; Realised hertability estimation; Design of selection experiments; Regression estimation; Maximum likelihood.

13

Variability of response to selection in genetic experiments

by

William G. Hill

BIOMETRICS *30*, 363–366
June 1974

362: Variability of Response to Selection in Genetic Experiments

WILLIAM G. HILL

Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN, Scotland

SUMMARY

Formulae are derived for predicting the variance of response to truncation selection. Allowance is made for variability in the selection differential, so these formulae differ from previous ones which were for the variance of response conditional on the selection differential applied. However, the magnitude of the genetic drift variance, comprising most of the variance in response, is not greatly affected by whether it is conditional or unconditional on the selection differential.

In some recent papers we have investigated the error structure for response to selection for quantitative traits in genetic experiments, in which account has been taken of genetic drift due to finite population size (Hill [1971; 1972a, b]). The analysis has been orientated towards estimates of parameters in the base population by means of the realized heritability, which is the regression of response on selection differential (Falconer [1960]). The relevant variances of mean performance or response each generation are then conditional on the selection differential applied, and all formulae in the previous papers are for conditional variances (usually implied rather than expressed explicitly in every formula). However, when examining selection results or making some predictions prior to starting an experiment, we may wish to estimate the variance between the response in several replicate lines in which selection has been practiced in the same way, e.g., by truncation selection of a fixed number of potential parents from a fixed number recorded. Whilst the expected selection differential will be the same for all replicates, there will be variance amongst them so that this unconditional variance of response will exceed that conditional on the selection differential. In this note we derive formulae for the unconditional responses. Other formulae for the variance of conditional response have been given by Prout [1962] for single generations, and incorrectly (see Hill [1971]) by Soller and Genizi [1967] for several generations. A result for the unconditional variance has been obtained by Baker [1972], but apparently incorrectly as we show here.

The genetic model is described by Hill [1971]. We consider an additive trait with phenotypic variance σ^2 and heritability h^2, so that h^2 is the regression of breeding value on phenotype, h is their correlation and they are bivariate normally distributed. In a monecious model M individuals are scored and N selected on their phenotype; the case of two sexes is discussed later. When variances of responses are expressed conditional on the selection differential it is not necessary to specify how selection of parents is carried out, so long as it is based only on the phenotypes of the M individuals and no other information. In the unconditional model the results are only of interest if the selection rules are more precisely specified. We shall assume that truncation selection is practiced and the highest ranking N individuals chosen. The proportion selected is $p = N/M$.

Consider one generation of selection. Let the mean breeding value of parents be μ, and the phenotypic mean of a random group of M progeny be $\bar{X}$. Let the phenotypic mean of the selected individuals by $\bar{Y}$ and their mean breeding value be $\bar{Z}$. From regression theory

$$\bar{Z} = \mu + h^2(\bar{Y} - \mu) + \bar{e}, \tag{1}$$

where $\bar{e}$ is the mean deviation of true breeding value from that predicted by regression, with $V(\bar{e}) = h^2(1 - h^2)\sigma^2/N$. The deviations $\bar{Y} - \mu$ and $\bar{e}$ are uncorrelated, hence

$$V(\bar{Z}) = h^4V(\bar{Y}) + h^2(1 - h^2)\sigma^2/N. \tag{2}$$

Now $V(\bar{Y})$ is the variance of the mean of the highest N order statistics from a sample of size M. Exact values for $V(\bar{Y})$ can be obtained only for small samples from Sarhan and Greenberg [1962]. However, Schaeffer *et al.* [1970] point out that for a given proportion selected the appropriate variance is very nearly inversely proportional to the number selected, N. Thus we can write

$$V(\bar{Y}) = k_p\sigma^2/N,$$

and approximate values of k_p are given by Schaeffer *et al.* [1970] (who denote this k_n). Some typical values are given in Table 1. Note that $k_{1-p} = 1 - p(1 - k_p)/(1 - p)$; for example $k_{0.8} = 0.866$, close to the value of 0.865 given by Schaeffer *et al.* Thus from (2)

$$V(\bar{Z}) = h^2\sigma^2[1 - (1 - k_p)h^2]/N. \tag{3}$$

The term k_p is absent from Baker's [1972] formula.

Let us now contrast the variance given by (3), which is not conditional on the selection differential, with the conditional variance. Following Hill [1971], rewrite (1) as

$$\bar{Z} = \mu + h^2(\bar{Y} - \bar{X}) + h^2(\bar{X} - \mu) + \bar{e}$$

and note that $\bar{Y} - \bar{X}$ is the observed selection differential. Thus

$$\begin{aligned} V(\bar{Z} \mid \bar{Y} - \bar{X}) &= h^4V(\bar{X}) + V(\bar{e}) \\ &= h^2\sigma^2[1 - (1 - p)h^2]/N. \end{aligned} \tag{4}$$

The difference between the conditional and unconditional variances is

$$V(\bar{Z}) - V(\bar{Z} \mid \bar{Y} - \bar{X}) = h^4\sigma^2(k_p - p)/N,$$

TABLE 1

COEFFICIENT OF ORDER STATISTICS (k_p) AND RATIO (R) OF UNCONDITIONAL TO CONDITIONAL DRIFT VARIANCES AS A FUNCTION OF PROPORTION SELECTED (p) AND HERITABILITY (h^2)

p	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
k_p	0.188	0.224	0.296	0.366	0.466	0.542	0.613	0.679
$R(h^2 = 0.1)$	1.020	1.023	1.027	1.029	1.029	1.026	1.023	1.019
$R(h^2 = 0.4)$	1.118	1.134	1.159	1.166	1.156	1.134	1.112	1.090

which is seen from the above table to approximate $0.2h^4\sigma^2/N$ for all selection intensities of interest, $0 < p < 0.5$. When no selection is practiced, $p = k_p = 1$, and $V(\bar{Z}) = V(\bar{Z} \mid \bar{Y} - \bar{X}) = h^2\sigma^2/N$, the usual formula for genetic drift in unselected populations.

The relative sizes of the conditional and unconditional variances can be expressed as

$$R = V(\bar{Z})/V(\bar{Z} \mid \bar{Y} - \bar{X})$$
$$= [1 - (1 - k_p)h^2]/[1 - (1 - p)h^2]$$

from (3) and (4). Values of R are given in Table 1 for $h^2 = 0.1$ and $h^2 = 0.4$. The size of R depends more on heritability than selection intensity, and does not depart far from unity, unless h^2 is high.

Extension of the results to several generations of selection is straightforward if we make the assumption that variances and heritabilities remain constant, for which the conditions have been discussed by Hill [1971]. Then equation (3) gives the drift variance, σ_d^2, per generation, since $\bar{Z}$ takes the place of μ in the following generation and the drift variance accumulates in proportion to generation number. There is an additional noncumulative source of error, due to estimation of the mean breeding value of the parents from the mean phenotype of the progeny. This variance depends on the distribution of family size, but is approximately equal to $V(\bar{X} - \mu) = \sigma^2/M$. In addition, we have to compute the covariance of generation means, which includes the genetic drift of the earlier generation, plus a noncumulative covariance between the mean phenotype of all recorded individuals and the breeding value of selected individuals that generation. This quantity is, from (1),

$$\text{cov}\,(\bar{X}, \bar{Z}) = h^2 \text{ cov}\,(\bar{X}, \bar{Y})$$
$$= h^2 V(\bar{Y}),$$

using the properties of order statistics (Sarhan and Greenberg [1962]); we get therefore

$$\text{cov}\,(\bar{X}, \bar{Z}) = h^2\sigma^2/M.$$

Letting $\bar{X}_i$ be the performance at generation i (so $\bar{X}_0$ corresponds to $\bar{X}$ above) we have the following unconditional variances

$$V(\bar{X}_i) = i\sigma_d^2 + \sigma_e^2 + h^2\sigma^2/M$$
$$\text{cov}\,(\bar{X}_i, \bar{X}_j) = i\sigma_d^2 + h^2\sigma^2/M, \qquad j > i$$

where $\sigma_d^2 = h^2\sigma^2[1 - (1 - k_p)h^2]/N$ and $\sigma_e^2 = (1 - h^2)\sigma^2/M$. The results expressed in this way correspond with those given for the same selection model, but with variances conditional on selection differentials, by Hill [1972b], and the only difference is a substitution of k_p for p in σ_d^2. If replicate lines are not contemporaneous, the variance between them will be increased by any environmental variance common to all individuals in a line at any generation. This common environmental variance does not accumulate and therefore should be added to σ_e^2 (Hill [1971]).

Extension of the results to divergent selection and to correlated traits is straightforward. In all cases k_p replaces p in formulae given in our earlier papers. With two sexes, where N_m males are selected from M_m, N_f females selected from M_f and $m = N_m/M_m$, $f = N_f/M_f$,

$$\sigma_d^2 = \tfrac{1}{4}h^2\sigma^2\{[1 - (1 - k_m)h^2]/N_m + [1 - (1 - k_f)h^2]/N_f\}$$
$$= h^2\sigma^2[1 - (1 - k_e)h^2]/N_e,$$

where $1/N_e = 1/4N_m + 1/4N_f$ and $k_e/N_e = k_m/4N_m + k_f/4N_f$. Similarly we replace M by its effective size, $M_e = 1/4M_m + 1/4M_f$, in σ_e^2.

We see that whilst there is a conceptual difference between the conditional and unconditional expressions for the sampling variance of the selection response, there is little difference in their magnitude (Table 1). Formally, however, each should be used where appropriate.

ACKNOWLEDGMENT

I wish to thank Dr. J. H. Louw for helpful discussion.

VARIABILITE DE LA REPONSE A LA SELECTION DANS DES EXPERIENCES GENETIQUES

RESUME

On déduit des formules pour prédire la variance d'une réponse à une sélection par troncation. On autorise une variabilité de la sélection différentielle, de telle sorte que les formules changent des précédentes qui ont été appliquées à la variance d'une réponse conditionnelle à la selection différentielle. Néanmoins la grandeur de la variance de dérive génétique, qui contient la plus grande part de la variance de la réponse, n'est pas affectée de facon importante, que la sélection différentielle soit conditionnelle ou inconditionnelle.

REFERENCES

Baker, R. J. [1971]. Theoretical variance of response to modified pedigree selection. *Can. J. Plant Sci. 51*, 463–8.

Falconer, D. S. [1960]. *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh.

Hill, W. G. [1971]. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics 27*, 293–311.

Hill, W. G. [1972a]. Estimation of realised heritabilities from selection experiments. I. Divergent selection. *Biometrics 28*, 747–65.

Hill, W. G. [1972b]. Estimation of realised heritabilities from selection experiments. II. Selection in one direction. *Biometrics 28*, 767–80.

Prout, T. [1962]. The error variance of the heritability estimate obtained from selection response. *Biometrics 18*, 404–7.

Sarhan, A. E. and Greenberg, B. G. [1962]. *Contributions to Order Statistics*. Wiley, New York.

Schaeffer, L. R., Van Vleck, L. D. and Velasco, J. A. [1970]. The use of order statistics with selected records. *Biometrics 26*, 854–9.

Soller, M. and Genizi, A. [1967]. Optimum experimental designs for realised heritability estimates. *Biometrics 23*, 361–5.

Received February 1973, *Revised July* 1973

Key Words: Genetics; Selection response.

Reprinted from
BIOMETRICS Copyright © 1974
THE BIOMETRIC SOCIETY, Vol. 30, No. 2, June 1974

14

Effective size of populations with overlapping generations

by

William G. Hill

Reprinted from THEORETICAL POPULATION BIOLOGY
Vol. 3, No. 3, September 1972
Printed in Belgium

Effective Size of Populations with Overlapping Generations

WILLIAM G. HILL

Institute of Animal Genetics, University of Edinburgh, Edinburgh, EH9 3JN

General formulae are derived for the effective sizes (numbers) of random mating populations of constant size and sex ratio with overlapping generations. They are found to equal the effective sizes of populations with discrete generations which have the same number of individuals entering the population each *generation* and the same variance of *lifetime* family number.

1. INTRODUCTION

Population and quantitative geneticists find that the concept of effective size or number, due initially to Wright, is useful for predicting inbreeding or random genetic drift. Most theoretical developments have been made with models of discrete generations, and these have been reviewed by those primarily concerned, namely, Wright (1969) and Crow and Kimura (1970). More recently, formulas for rates of inbreeding or effective size have been developed for specific models with overlapping generations by Moran (1962), Kimura and Crow (1963), Nei and Imaizumi (1966), Felsenstein (1969, 1971), Turner and Young (1969) and both A. Robertson (private communication) and J. W. James (private communication) have developed relevant, but unpublished formulas. Felsenstein (1971) develops rigorous methods, and using Moran's specific model of random births and deaths, he finds that the formula of Kimura and Crow (1963) is incorrect, and that of Nei and Imaizumi (1966) is, at best, vague. Crow and Kimura (1971) have retracted their earlier formula, and given a new one in terms of age-specific birth and death rates. In none of these formulas can real differences in fertility, i.e., unequal expectation of progeny numbers among survivors, be incorporated. While Giesel (1969) claims to include these effects, it is not apparent from his formulas how this should be done. However, Nei (1970) and Crow and Kimura (1971) use the discrete generation analogy to obtain equations for effective size in the presence of fertility differences, but give no proof for the overlapping model.

In this paper a more general result for the variance effective number is derived in which the variance of family size can be specified. However, the population is assumed to be maintained with a constant size and age distribution, and with

these and other small restrictions on the model we find that the final result is a simple extension of that for discrete generations. This particular study was initiated to find the effective size of control populations for selection experiments and breeding programmes (see review by Hill, 1972) in which these criteria of the model can be met. For comparison of alternative methods of maintaining controls we shall find it useful not only to consider the effective population size N_e, which is the size of an idealised population leading to the same variance of gene frequency change per generation, but also an annual effective size N_y, which is the size of an idealised population with a generation interval of one year leading to the same variance of gene frequency change per year. Thus, if the population has generation interval L, and the increment in variance is constant per year, $N_y = LN_e$.

The problem is tackled in three stages: first, by considering a general stochastic process, then putting this into a genetic context and invoking variable family size in a haploid model, and finally considering diploid models.

2. Basic Model

Let us assume that parents may be of age 1, 2,..., k years (or other time units) and that the gene frequency of individuals born at year t is q_t. Further, let the proportion of genes derived from individuals of age i have expectation p_i, where the p_i include both survival to, and fertility at, age i, and $\sum_{i=1}^{k} p_i = 1$. Since the population is finite, there is a sampling error or drift d_t, associated with the individuals born in year t, due to chance deviations in viability and fertility. Thus

$$q_t = \sum_{i=1}^{k} p_i q_{t-i} + d_t , \tag{1}$$

where $E(d_t) = 0$.

We now consider some late generation and express its gene frequency in terms of the gene frequency in successively earlier generations and the drift in the intervening generations. Replacing q_{t-1} by q_{t-2},..., q_{t-k-1} and d_{t-1} in (1), we obtain

$$q_t = \sum_{i=2}^{k} p_i q_{t-i} + p_1 \sum_{i=1}^{k} p_i q_{t-i-1} + d_t + p_1 d_{t-1}$$

$$= \sum_{i=1}^{k-1} (p_1 p_i + p_{i+1}) q_{t-1-i} + p_1 p_k q_{t-1-k} + d_t + p_1 d_{t-1} .$$

Now q_{t-2} is replaced by $q_{t-3},\ldots, q_{t-2-k}, d_{t-2}$, and then q_{t-3} by $q_{t-4},\ldots, q_{t-3-k}, d_{t-3}$ and so on in succession. Thus for any T, $0 \leqslant T \leqslant t$, we can write, in succession,

$$q_t = \sum_{i=1}^{k} x_{Ti} q_{t-T-i} + \sum_{j=0}^{T} y_j d_{t-j}, \tag{2}$$

where x_{Ti} and y_j are appropriate constants to be found. For example, $x_{0i} = p_i$, $x_{12} = p_1 p_2 + p_3$, $y_0 = 1$, $y_1 = p_1$. Using (1) to replace q_{t-T-1} by $q_{t-T-2},\ldots, q_{t-T-1-k}, d_{t-T-1}$ in (2), we get

$$q_t = \sum_{i=1}^{k-1} (p_i x_{T1} + x_{T,i+1}) q_{t-T-1-i} + p_k x_{T1} q_{t-T-1-k} + x_{T1} d_{t-T-1} + \sum_{j=0}^{T} y_j d_{t-j}. \tag{3}$$

But, in (3), we have now obtained the coefficients $x_{T+1,i}$ and y_{T+1} as follows:

$$\begin{aligned} x_{T+1,i} &= p_i x_{T1} + x_{T,i+1}, \qquad i = 1,\ldots, k-1, \\ x_{T+1,k} &= p_k x_{T1}, \\ y_{T+1} &= x_{T1}. \end{aligned} \tag{4}$$

Denoting by $\mathbf{x}_T$ the column vector with elements x_{Ti}, $i = 1,\ldots, k$ and $\mathbf{M}$ a square matrix of dimension k specifying the recurrence relations, we have

$$\mathbf{x}_{T+1} = \begin{pmatrix} p_1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ p_2 & 0 & 1 & 0 & \cdots & 0 & 0 \\ p_3 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ p_{k-1} & 0 & 0 & 0 & \cdots & 0 & 1 \\ p_k & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \mathbf{x}_T = \mathbf{M}\mathbf{x}_T.$$

The matrix $\mathbf{M}$ is a special case of the type discussed by Leslie (1945). It has a single eigenvalue of 1, and all others are negative or complex with absolute value less than unity. Thus if $\mathbf{x}$ is the eigenvector associated with the unit eigenvalue then $\mathbf{x}_T$ approaches $\mathbf{x}$ as T increases. The solution of $\mathbf{x} = \mathbf{M}\mathbf{x}$ gives this eigenvector, which is found to be

$$\begin{aligned} x_1 &= c, \\ x_i &= c\left(1 - \sum_{j=1}^{i-1} p_j\right), \qquad i > 1, \end{aligned} \tag{5}$$

where c is an arbitrary constant. But since $x_{0i} = p_i$ and $\sum_{i=1}^{k} p_i = 1$, it is clear from (4) that $\sum_{i=1}^{k} x_{Ti} = \sum_{i=1}^{k} x_i = 1$. Imposing this restraint on (5) we obtain $c = 1/\sum_i i p_i$. But the generation interval L is given by $L = \sum_i i p_i$; therefore $x_1 = 1/L$. A derivation similar to that above has been provided by Goodman (1969).

From (4) we see that the coefficients y_j of d_{t-j} also approach $1/L$ as we consider generations increasingly earlier than t. If initially $q_0 = q_{-1} = \cdots = q_{1-k}$, then from (2),

$$q_t = q_0 + \sum_{j=0}^{t-1} y_j d_{t-j} ,$$

or, rearranging,

$$q_t = q_0 + \frac{1}{L} \sum_{j=1}^{t} d_j + \sum_{j=1}^{t} \left(y_{t-j} - \frac{1}{L} \right) d_j . \tag{6}$$

In the simplest model of this process, we could assume that the d_j are uncorrelated with constant variance σ_d^2. Then as t increases, the drift per year (time period) $V(q_t)/t$ approaches σ_d^2/L^2, or that per generation $V(q_t)/(t/L)$ approaches σ_d^2/L. These simple assumptions can not be made in the genetic context, for correlation of the d_j will result from premature death, for example, or high fertility over several years of breeding individuals.

3. Haploids

Consider a haploid model in which N individuals are born every year. Let q_{Tl} be the frequency of one of two alternative neutral alleles of individual l born in year T, and let it have $n_{T+i,T,l}$ progeny (or haploid replicas) in year $T + i$. Thus $q_{Tl} = 0$ or 1, and

$$q_T = \frac{1}{N} \sum_{l=1}^{N} q_{Tl}$$

$$= \frac{1}{N} \sum_{i=1}^{k} \sum_{l=1}^{N} q_{T-i,l} n_{T,T-i,l} , \tag{7}$$

and $\sum_i \sum_l n_{T,T-i,l} = N$. From (1) and (7),

$$d_T = \frac{1}{N} \sum_i \sum_l (n_{T,T-i,l} - p_i) q_{T-i,l} . \tag{8}$$

We now use (6), with T replacing j, and make the same assumptions about the base population. We are primarily concerned with the effects on variance of

some intermediate generation, without the complications involved at the ends of the process. Thus we concentrate on those individuals which have all their parents born after year 0, and have all their progeny by year t. From (6) and (8),

$$q_t = q_0 + \frac{1}{NL} \sum_{T=k+1}^{t-k} \sum_i \sum_l (n_{T+i,T,l} - p_i)\, q_{T,l} + R, \tag{9}$$

where R has terms derived both from the departure of y_{t-T} from $1/L$ and from the first few and last few generations. Let n_{Tl} be the total number of progeny got by the specified individual in its lifetime, i.e., $n_{Tl} = \sum_i n_{T+i,T,l}$, and let n_T be the mean number of individuals born in year T, i.e., $n_T = 1/N \sum_l n_{Tl}$. Equation (9) becomes

$$\begin{aligned} q_t - q_0 &= \frac{1}{NL} \sum_{T=k+1}^{t-k} \sum_l (n_{Tl} - 1)\, q_{Tl} + R \\ &= \frac{1}{NL} \sum_{T=k+1}^{t-k} \left[\sum_l (n_{Tl} - n_T)(q_{Tl} - q_T) + N(n_T - 1)\, q_T \right] + R. \end{aligned} \tag{10}$$

Now we make the reasonable assumptions that there are no covariances of gene frequency with family size, nor of deviations about their mean of gene frequency or family sizes between groups born at different times; thus,

$$\begin{aligned} \operatorname{cov}[(n_{Tl} - n_T)(q_{Tl} - q_T)] &= 0, \\ \operatorname{cov}[(q_{Tl} - q_T)(q_{T'l'} - q_{T'})] &= 0, \qquad T \neq T', \\ \operatorname{cov}[(n_{Tl} - n_T)(n_{T'l'} - n_{T'})] &= 0, \qquad T \neq T'. \end{aligned}$$

We make the further, but more limiting, restriction that the total number of progeny got by a group born in any particular year is exactly N, so that $n_T = 1$. The second term in (10) then vanishes and we bypass the difficult problem of finding the covariance structure of q_T. Defining the variance of observed lifetime family size of individuals born in year T as

$$\sigma^2_{nT} = \sum_{l=1}^{N} (n_{Tl} - n_T)^2/N$$

and noting that

$$V(q_{Tl} - q_T) = -(N-1) \operatorname{cov}(q_{Tl} - q_T\,,\, q_{Tl'} - q_T) = q_T(1 - q_T)$$

for $l \neq l'$, we have from (10)

$$V(q_t) = \frac{1}{(N-1)L^2} \sum_{T=k+1}^{t-k} \sigma_{nT}^2 q_T(1-q_T) + V(R)$$

$$+ \operatorname{cov}\left(\frac{1}{NL} \sum_{T=k+1}^{t-k} \sum_{l} (n_{Tl} - n_T)(q_{Tl} - q_T), R\right). \qquad (11)$$

Consider the contribution from individuals born in some year T, $k+1 \leqslant T \leqslant t-k$, sufficiently early that the drift coefficient is close to $1/L$. Then there is no contribution of these individuals to $V(R)$ or to the covariance term in (11), and the increment $V_T(q_t)$ for which they are responsible in V_T is $V_T(q_t) = \sigma_{nT}^2 q_T(1-q_T)/(N-1)L^2$. The variance of gene frequency change in an idealised haploid population with a generation interval of one year is $q_T(1-q_T)/N_y$, where N_y is the annual effective size, so that

$$N_y = (N-1)L^2/\sigma_{nT}^2 . \qquad (12)$$

If the population is sufficiently large so that there is little change in q_T over a period of one generation, and $\sigma_{nT}^2 = \sigma_n^2$ remains constant over years, then the increment in variance per generation is $\sigma_n^2 q_T(1-q_T)/(N-1)L$ and

$$\begin{aligned} N_e &= N_y/L \\ &= (N-1)L/\sigma_n^2 . \end{aligned} \qquad (13)$$

With a random sampling of family sizes (i.e., multinomial distribution, or Poisson distribution restrained by a fixed total number), $\sigma_n^2 = 1 - 1/N$ and $N_e = NL$, the number of individuals entering the population each generation. This result agrees with that of Felsenstein (1971), and can also be shown to give the correct effective size for the Moran model. For haploids with this model, individuals have an exponential distribution of lifetime, and on death are replaced by a duplicate of the recently dying or any other individual, each with equal probability. From formulas given by Moran (1962), it can be shown that $\sigma_n^2 = 2(1 - 1/N^*)$, where there are N^* individuals in the population. Thus $N_e = N^*/2 = NL/2$ as Moran shows.

4. Monecious Diploids

Little change in formulation is required to enable extension of these results to a monecious diploid model; the methods are essentially an extension of those of Crow and Kimura (1970), and a rigorous proof is not given here. If there is

random mating of all individuals alive in the population at any time, each progeny is formed from the random sampling of 2 genes; thus if N are born in one year a total of $2N$ genes are sampled. Thus in (8) we replace N by $2N$, and $q_{T-t,t}$ now takes the values 0, $\frac{1}{2}$ or 1. But, in addition, sampling occurs from segregation within heterozygous parents (having $q_{T-t,t} = \frac{1}{2}$), which is independent of the sampling occurring between parents. The expected increment in the variance of drift d_t from this source is therefore $(1/4N)x$ expected proportion of heterozygotes, and such drift is independent in successive time periods. The proportion of heterozygotes among individuals born in year T, given that they have mean gene frequency q_T, is $2q_T(1 - q_T)[2N/(2N - 1)]$. The small departure from Hardy–Weinberg frequencies is derived by Crow and Kimura (1970) for discrete populations, and is based on the sampling without replacement of genes, conditional on q_T.

Combining the sampling between parents and within heterozygotes, we obtain

$$N_e = (4N - 2)L/(\sigma_n^2 + 2), \tag{14}$$

where N diploid individuals enter the population per year and σ_n^2 is the variance of lifetime family size or, strictly, the variance of the number of gametes represented in the next generation. With equal family sizes $N_e = (2N - 1)L$, and with a multinomial distribution of sizes, with mean 2, $\sigma_n^2 = 2(1 - 1/N)$ and $N_e = NL$.

Since the gene frequencies of individuals born in different years are not exactly the same, there may be a slight excess of heterozygotes above that predicted. In addition, the other assumption of exactly equal numbers of progeny from individuals born in each year is unlikely to be realised in practice. However, if the population is sufficiently large so that there are only small differences in mean frequencies in successive years, this departure from assumption can have little effect on the drift variance. But the formulas are then not precise to terms of order $1/N$. Thus, an adequate approximation to (14) for the effective size is

$$N_e = 4NL/(2 + \sigma_n^2), \tag{15}$$

which is the same as that of a population having discrete generations, with the same number of individuals entering the population per *generation*, and the same variance of *lifetime* family size.

An example will now be given to show how (15) can be evaluated for models of age specific birth and death rates. A discrete time model could be used, but the formulation is simpler in the continuous case, although some biological problems such as the availability of mates are ignored. Results are stated without proof; for reference, a text on stochastic processes should be consulted (e.g., Karlin, 1966). Let l_y be the probability of survival to age y, and let $b_y\,dy$ be

the expected number of births of a survivor in the age interval y to $y + dy$. The probability density of the age at death is then $f(y) = -dl_y/dy$. The expected number of progeny given by an individual dying at age y is $n_y = \int_0^y b_w\, dw$, and the average family size is $\bar{n} = E(n_y)$, where expectation here and elsewhere is taken over $f(y)$. If the population is of constant size, $\bar{n} = 2$ for the monecious diploid model. Assuming that births occur randomly, the distribution of the number of progeny, conditional on the age at death will be Poisson, with conditional variance n_y. The unconditional variance of family size σ_n^2 is comprised of two parts: that from variance in age at death, and that from variance in family size conditional on age at death. Thus,

$$\begin{aligned}\sigma_n^2 &= E(n_y) + E(n_y^2) - E^2(n_y) \\ &= E(n_y^2) - 2,\end{aligned}$$

if $\bar{n} = 2$. Thus, from (15), $N_e = 4NL/E(n_y^2)$. The generation interval is given by $L = \int_0^\infty y l_y b_y\, dy/\bar{n}$ and the mean age of individuals by $A = E(y)$.

We consider the example given by Crow and Kimura (1971), but with the birth rate doubled since they were concerned with haploids. We let $l_y = e^{-y/N}$ and $b_y = 4N^{-1}e^{-y/N}$. Thus, $f(y) = N^{-1}e^{-y/N}$, $n_y = 4(1 - e^{-y/N})$, $E(n_y) = \bar{n} = 2$ and the population size is constant; $L = N/2$, $A = N$, and finally, $\sigma_n^2 = 10/3$. If we assume that the total size of the population is N^*, on average N^*/N individuals enter the population per time unit. Thus, from (15), $N_e = 3N^*/8$. This ratio of 3/8 of effective to actual size is also obtained for the analogous haploid model, using either the formulation of this paper, or that of Felsenstein (1971). However the Crow-Kimura approximation gives a ratio of 1/3 (Crow and Kimura, 1971).

In the above formulation for the continuous model, no restriction is imposed on family sizes to ensure that the population size remains constant, nor that exactly one individual is born per time unit. This does not seem to be a serious restriction, for if the same approach is used for a continuous version of the Moran model in which $l_y = e^{-y/N}$ and $b_y = dy/N$, then $\sigma_n^2 = 2$, rather than the correct value of $2(1 - 1/N)$ noted in the previous section.

Other assumptions are also implicit in the approach used in this paper. The parameters of our model have been specified in terms of the distribution of number of newborn progeny among newborn individuals, many of whom have no progeny if they die prior to reproductive age. The assumption is being made that there are no familial correlations, either in viability or fertility. If there are, the drift variance is smaller and the effective size larger than that predicted. This could be measured in terms of the variance of the numer of grandprogeny of each individual, which would exceed the variance expected from two generations of independent sampling of progeny number. A familial correlation of viability to adulthood can be taken into account by specifying

the number of adult individuals (N_a) entering the population each generation, and the variance of the number of adult progeny per adult individual (σ_a^2). The formulation used earlier still applies directly, and the effective size from (15) becomes

$$N_e = 4N_aL/(2 + \sigma_a^2), \tag{16}$$

assuming the generation interval is unchanged. Formulas essentially the same as (16) have been suggested by Nei (1970) and Crow and Kimura (1971).

Usually information is available on the variance of the number of newborn progeny among adults (say σ_s^2). If there is no familial correlation of survival, it can be shown that

$$\sigma_a^2 = 2d + (1 - d)^2 \sigma_s^2,$$

where d is the proportion which die prior to reaching adulthood and the population size is static, so that the mean family size at birth is $2/(1 - d)$ (from Crow and Morton, 1955). This formula for σ_a^2 can be substituted into (16). Further discussions of the problems of familial variation in fertility, at least in discrete generation models, are given by Crow and Morton (1955) and Nei and Murata (1966).

5. Separate Sexes

When sexes are separate we need to specify variances and covariances of numbers of male and female progeny among male and female parents. As before, individuals are assumed to mate at random. It is possible to define the effective size either in terms of numbers of adults or newborns; we use the former so that familial correlation of viability is adequately included.

Each year, let us assume M males and F females reach adulthood (or in the animal breeding context are taken for breeding). The mean and variances of the number of progeny reaching adulthood are as follows:

Pathway for gametes		*Mean*	*Variance*
Male parents having male progeny		1	σ^2_{mm}
male	female	F/M	σ^2_{mf}
female	male	M/F	σ^2_{fm}
female	female	1	σ^2_{ff}

Also, let the covariance of the number of male and female progeny from each male parent be cov(mm, mf) and from each female be cov(fm, ff). These covariance terms are defined by Latter (1959) for discrete generation models but have been ignored by Crow and Kimura (1970). Where most of the variance

in family size is caused by differential mortality, these covariances are likely to be of the same order as the variances and should not be omitted, whether generations are discrete or overlapping. Omitting terms of order $1/M$ or $1/F$ relative to 1, it can be shown by extending the results for the monecious populations that

$$\frac{1}{N_e} = \frac{1}{16ML}\left[2 + \sigma^2_{mm} + 2\left(\frac{M}{F}\right)\mathrm{cov}(mm, mf) + \left(\frac{M}{F}\right)^2 \sigma^2_{mf}\right]$$
$$+ \frac{1}{16FL}\left[2 + \left(\frac{F}{M}\right)^2 \sigma^2_{fm} + 2\left(\frac{F}{M}\right)\mathrm{cov}(fm, ff) + \sigma^2_{ff}\right], \qquad (17)$$

where L is again the generation interval (average age of parents along the 4 pathways for gametes). Again, the formula is that for discrete generations with the same numbers entering per generation and the same variance of lifetime family size. If there is no differential viability or fertility, family sizes take the Poisson distribution and there is no covariance of numbers of male and female progeny. Then (17) reduces to

$$1/N_e = L/N_y = (1/4ML) + (1/4NL). \qquad (18)$$

In the context of control populations we see that the effective size of the population is increased by minimising the variance of family sizes, but it is immaterial whether a sire has all his progeny when 2 years of age, or an equal proportion when 1, 2 or 3 years old, so long as the total number is the same. It is also possible to show that the effective size (N_e) and annual effective size (N_y) is increased by rapidly replacing males if females breed for many years (Turner and Young, 1969; Hill, 1972).

6. Concluding Remarks

The main result we have obtained is a proof that the formulas for effective population size in random mating populations of constant size which are appropriate for discrete generations can be applied directly to overlapping generations. This does not mean that effective sizes are not likely to be different in the two cases. For example, a Poisson or multinomial distribution of family size may be a reasonable hypothesis with discrete generations, but with overlapping generations will predict too low a variance of family size if some animals die before the end of their reproductive life. Thus simple formulas such as (18) in which family sizes are assumed to be Poisson distributed should be used with caution. The other important difference between the two kinds of population is that with discrete generations the drift variance is proportional to the computed

value of $1/N_e$ from the outset, whereas the variance approaches this value asymptotically if generations overlap.

The analysis has been restricted to the prediction of the variance effective size. However, in random mating populations of constant size, the variance and inbreeding effective sizes must be the same, whether or not generations overlap, since the total drift variance is proportional to the increase in homozygosity. Formal proofs are available for discrete generations (Kimura and Crow, 1963) and for some models of overlapping generations (Felsenstein, 1971; Crow and Kimura, 1971).

Several loose ends remain. When considering diploid models, no distinction was made between random mating among all adults, or among only those of the same age. In no case was the effect of removing the assumption of equal numbers of progeny in the lifetime of each cohort properly considered. Nevertheless, some reasons were given for suggesting that neither effect has much influence on effective population size.

Acknowledgments

I wish to thank Dr. J. F. Crow and Dr. J. Felsenstein for copies of their manuscripts before publication and, together with a referee, for their helpful comments on this paper.

References

Crow, J. F. and Kimura, M. 1970. "An Introduction to Population Genetics Theory," Harper and Row, New York.

Crow, J. F. and Kimura, M. 1971. The effective number of a population with overlapping generations: A correction and further discussion, *Amer. J. Hum. Genet.* **24**, 1–10.

Crow, J. F. and Morton, N. E. 1955. Measurement of gene frequency drift in small populations, *Evolution* **9**, 202–214.

Felsenstein, J. 1969. The effective size of a population with overlapping generations, *Genetics* **61**, s18.

Felsenstein, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations, *Genetics* **68**.

Giesel, J. T. 1969. Inbreeding in a stationary, stable population as a function of age and fecundity distribution, *Genetics* **61**, s21.

Goodman, L. A. 1969. The analysis of population growth when the birth and death rates depend upon several factors, *Biometrics* **25**, 659–691.

Hill, W. G. 1972. Estimation of genetic change, I: General theory and design of control populations, *Anim. Breed. Abstr.* **40**, 1-15.

Karlin, S. 1966. "A First Course in Stochastic Precesses," Academic Press, New York.

Kimura, M. and Crow, J. F. 1963. The measurement of effective population numbers, *Evolution* **17**, 279–288.

Latter, B. D. H. 1959. Genetic sampling in a random mating population of constant size and sex ratio, *Aust. J. Biol. Sci.* **12**, 500–505.

LESLIE, P. H. 1945. On the use of matrices in certain population mathematics, *Biometrika* **33**, 213–245.

MORAN, P. A. P. 1962. "The Statistical Processes of Evolutionary Theory," Clarendon Press, Oxford.

NEI, M. 1970. Effective size of human populations, *Amer. J. Hum. Genet.* **22**, 694–695.

NEI, M. AND MURATA, M. 1966. Effective population size when fertility is inherited, *Genet. Res.* **8**, 257–260.

NEI, M. AND IMAIZUMI, Y. 1966. Genetic structure of human populations, II: Differentiation of blood group gene frequencies among isolated populations, *Heredity* **21**, 183–190, 344.

TURNER, H. N. AND YOUNG, S. S. Y. 1969. "Quantitative Genetics and Sheep Breeding," Macmillan, Melbourne.

WRIGHT, S. 1969. "Evolution and the Genetics of Populations, 2: The Theory of Gene Frequencies," University of Chicago Press, Chicago, Ill.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.

15

Estimation of genetic change

I. General theory and design of control populations

by

William G. Hill

Commonwealth Bureau of Animal Breeding and Genetics

REPRINT No. 92

Reprinted from ANIMAL BREEDING ABSTRACTS, VOL 40. No. 1, pp., 1–15, MARCH 1972

ESTIMATION OF GENETIC CHANGE. I. GENERAL THEORY AND DESIGN OF CONTROL POPULATIONS

W. G. HILL

Institute of Animal Genetics, Edinburgh

The separation of observed change into its environmental and genetic components is an important part of the analysis of selection experiments or breeding programmes. It is rarely possible to conduct experiments in uniform conditions over periods of several generations, so that changes in performance of a selected population may reflect, in part, some environmental change. The problem was recognised by Lerner in 1950, who discussed the confounding of effects and suggested ways of separating them.

Several methods of estimating genetic changes have now been devised. Many of these have been considered in recent reviews by Dickerson (1969) and Lindström (1969), but to differing depths. One of the most common methods is to utilise unselected control populations, but neither the design nor the properties of control populations observed in applied breeding or experimental situations have been analysed fully recently, although earlier discussions are available (King *et al.*, 1959; Gowe *et al.*, 1959). The present review is therefore devoted primarily to control populations, although in part I several alternative methods of estimating change are described and their merits and efficiencies compared, but not in detail. Attention is given to planned methods rather than to methods based on field records which have not been collected specifically to enable response to be measured. Then, *theoretical* aspects of the design and possible limitations of control populations are considered in greater depth. In part II, a review will be undertaken of the results of *experimental* checks and analyses of control population stability, together with other experiments in which controls have merely been carried alongside selected lines. In view of some recent discussion by Clayton (1968) and Dickerson (1968) on the reliability of a widely used control population in poultry, this analysis may be timely.

Methods of Estimating Genetic Change

Constant environment

There are some special cases, particularly in laboratory animals, where the environment can be maintained sufficiently constant for many generations so that no fluctuation in mean performance of, say, a large unselected population can be observed. In such situations, genetic change can be estimated directly from phenotypic change. For example, this seems to be possible in *Drosophila melanogaster* populations, maintained at nearly constant temperature, for a trait such as bristle number, which is particularly insensitive to environmental variation (Clayton *et al.*, 1957). But egg production in *D. pseudoobscura* can show marked fluctuations over a period of generations, even when the flies are maintained at constant temperature (Kojima and Kelleher, 1963). Thus, unless there is prior evidence for stability of the particular trait measured on a species in some specified environment, a simple measure of response using phenotypic change cannot be used. Attempts have to be made to compare different genotypes at the same time in the same environment.

Comparison of alternative selection schemes

A measure of response to selection which is not confounded by environmental effects can be obtained by practising selection in opposite directions in two contemporaneous lines at the same location. This technique of divergent selection has been used most often with mice (Roberts, 1965). However, no precise estimate of asymmetry of response is possible so that certain checks of results against predictions cannot be made. If a selection experiment is planned solely for estimation of genetic parameters, such as realised heritability as determined by the regression of response on selection differential (Falconer, 1960), divergent selection can be shown to be the most efficient design, for no facilities are wasted on control populations (Hill, 1972*b*). Similarly, when the only objective in an experiment is to enable comparisons of alternative selection schemes for improving the same trait, the differences in response can be estimated without recourse to a control (*e.g.* Falconer and Latyszewski, 1952; Bell *et al.*, 1955), but the magnitudes of the actual responses from any particular scheme can not be estimated accurately.

Replication of the same genetic material in successive generations

If it is possible to replicate the same set of genotypes in successive generations in some population, it can be used as a standard for comparison with a selected population. A change in the difference between the performances of the two populations when maintained in the same environment at the same time is then an estimate of genetic change in the selected population. The main criterion by which any such method should be judged is the precision of the estimate of response which is obtained. The accuracy will be reduced if the supposed standard undergoes genetic change itself, if it and the selected population react differently to environmental trend or fluctuation, or if few animals are measured. An assumption common to all the methods is that there is no accumulation of mutations which might cause bias or random error.

There are several ways in which replication of the same genetic material in successive generations can be approached or achieved. These methods will be considered in turn; to some extent the classifications are arbitrary, and there is considerable overlap among them.

Genotype storage. Genetic change in the control is eliminated by storage of the base material and sampling it whenever comparisons with the test population have to be made. This is possible in most plant species, where seed can be stored at low moisture content and low temperature for many years (James, 1961), or the plant can be reproduced vegetatively (Larson, 1961). Among the animal species, *Tribolium castaneum* can be maintained for at least 9 generations (*i.e.* months) by storage of adults at low temperature (Bray *et al.*, 1962). In this case, progeny are reared from the stored adults when required. In species with a long life span, the adult individuals themselves can be retained for the equivalent of several generations, again, for example in *Tribolium* (Bray *et al.*, 1962). For these methods to be unbiased there should be no correlation of survival, whether in storage or as adults, with any quantitative traits in which the control is to be compared with test material. By ensuring that there is little mortality, or in the case of seeds, loss of germination, selection can be minimised. An important requirement is that there should be no effect of ageing on the performance of the seed or on the progeny of stored individuals, whichever are used in the test. Unfortunately, storage of complete genotypes is not feasible with most animal species; should long-term ovum storage become possible in mammals, this might prove a useful technique.

Chromosome storage. With *Drosophila melanogaster* it is possible to sample a set of individual chromosomes from a population and maintain these separately for many generations balanced against marked inversion chromosomes. A sample of genotypes could be constituted from the chromosomes and used as the base population for a selection experiment, and a new sample constituted whenever an estimate of response was required. This technique would be very laborious for more than one particular chromosome (*e.g.* III), and is not feasible in other species.

Gamete storage. In cattle, gametes can be stored for long periods as deep-frozen semen. If every few years, depending on the generation length, females are inseminated with this semen and their progeny reared alongside progeny of sires in current use, an estimate of one half of the genetic change in the population can be obtained (Dickerson, 1960, 1969). However, only the additive component of change is estimated without bias. If a population is becoming inbred during selection and the progeny reared from the long-term stored semen are less inbred, then an underestimate of the change in genotypic value in the selected population would be obtained from the test comparison. The females which are bred with stored semen should have a similar age distribution to those mated to contemporary bulls, and should be chosen at random or matched for performance. The difference between the progeny performance of the old and contemporary bulls can then be estimated within age-of-dam classes.

Various semen storage plans are possible in cattle: semen can be retained from a random sample of young bulls, but then the variance of the estimate of over-all change includes part of the variance between bulls' breeding values; alternatively semen can be retained from bulls which have had an accurate progeny test, so that their merit at the start of the programme is known accurately. If high-ranking bulls are used initially, their later use will not depress performance and thus increase testing costs too greatly, even if real responses are being made. However, since these bulls have been selected on the basis of their progeny test, their initial merit must be judged on the basis of records obtained following the selection, or their progeny test will apparently regress (Dickerson, 1969). The semen storage method does not measure change in the population as a whole: the direct comparison is between bulls born in different years, and an implicit assumption is made that the population of cows is responding similarly. Thus, the estimate could be incorrect unless the population is closed to outside breeding stock. A small bias can be introduced if the selection of young bulls for future progeny testing improves with time, say by superior identification of good bull mothers. These errors will become smaller, at

least for estimates of yearly change, as the period over which comparisons are made is lengthened, for an increasing differential between cow and bull breeding values is unlikely.

Inbred lines. Essentially equivalent to the storage of individuals is the maintenance of highly inbred lines, which, generally in the form of inter-crosses, can be used as controls (Bell *et al.*, 1955; Rahnefeld *et al.*, 1963). However, a few lines represent only a small sample of genotypes, as do the crosses among them, and the lines are a special sample since they must have originally survived an intense inbreeding process. It is therefore possible that they will react in a unique way to environmental change. Highly inbred lines have not been obtained in the large animal species, nor in some much smaller ones, such as the Japanese quail (Sittmann *et al.*, 1966), and unless the lines are almost isogenic, steps may have to be taken to minimise further drift or natural selection within them. The other potential hazard is the accidental loss of one or more of the lines, although it may be possible to correct the data accordingly, with some reduction in precision.

Control populations. As an alternative to storing complete genotypes or keeping inbred lines in which no genetic change should occur, a segregating population can be maintained in which attempts are made to minimise the genetic change from selection or random drift. Widespread use has been made of such segregating control populations, commonly referred to simply as " control populations ", in both experimental and applied breeding situations, and there have been some discussions of their design, particularly for poultry (*e.g.* King *et al.*, 1959; Gowe *et al.*, 1959). The theory underlying the design of control populations is discussed in some detail later in this paper.

Repeat mating designs. A formalised method of genotype storage for one year, or generation, is the repeat mating design proposed primarily for poultry by Goodwin *et al.* (1955, 1960) and further elaborated by Dickerson (1961, 1965, 1969). These papers should be consulted for details of the somewhat complicated design, and, together with that of Giesbrecht and Kempthorne (1965) for a discussion of the associated sampling errors. They consider a primary population in which every year selection is practised and a new generation reared and estimates of change in this population are required. As a basis, X_{ij} denotes the population, or its mean performance, hatched in year i from pullet dams which have undergone j generations of selection, and Y_{ij} denotes birds hatched from older dams. The basic idea is to repeat matings of selected individuals from the primary population, X_{00}, in two successive years. Selection is practised among those hatched in the first year, X_{11}, to give a group, X_{22}, the primary population after two selections, and these are contemporaries of a group, Y_{21}, hatched from the repeat matings. Although $X_{22}-Y_{21}$ gives an estimate of genetic change from the second selection, it is confounded with maternal age effects, so further refinement is necessary. A set of matings is now made using the same males as in X_{11} together with females from Y_{21} which are full sisters of those in X_{11}, similarly selected. The progeny of Y_{21}, namely X_{32}, have had the same selection history and have dams of the same age as X_{22}, but are hatched one year later. Thus $X_{32}-X_{22}$ is an estimate of environmental change from year 2 to year 3, and since X_{32} are contemporaries of the progeny, X_{33}, of the next selection in the primary population, the difference $X_{33}-X_{32}$ is an estimate of genetic change from this last selection, free of maternal age effects.

A single repeat mating population, whether or not it is under selection, can be used to estimate environmental change and therefore be employed as a control for other selected populations maintained in the same environment.

Although not an essential part of the repeat mating design as such, Dickerson (1961, 1965) included progeny obtained from matings of randomly chosen individuals from the primary population, and was thus able to estimate the effects of relaxation of selection. A balance sheet of potential genetic response, recombination loss or natural selection on relaxation, and net genetic change could be constructed. An alternative control to the repeat mating method could be used and this partition still be made.

Instead of repeated complete matings, where it is difficult to remove maternal age effects, sires alone can be used in two or more consecutive years and compared with progeny of sires born the following year. Although similar in principle to the method of semen storage described above, a specific design has been suggested by Hickman and Freeman (1969), which was originally planned as an internal control in a small dairy cattle population, but the authors point out that the method could be used in other species. Young, unproven bulls enter the herd every year, and for a period of two years are mated to cows of all ages. Thus, each year contemporary comparisons, with maternal age effects eliminated, can be obtained of progeny of bulls born in successive years. To some extent, some loss of genetic response must be associated with structuring the herd to permit control comparisons, but this may be small for this specific design. The method has the particular advantage that few or no facilities are devoted to estimating the change. If young bulls were used for longer periods to allow more precise estimates of change, then greater sacrifices in response would have to be made. In an

earlier report, Hickman (1958) suggested a method in which groups of cows were mated to the same group of tested bulls in successive years, but maternal age effects would be confounded with the estimate of change.

These methods of repeat matings have a potential advantage over other methods utilising genotype or gamete storage in that the control and selected populations are highly related, so that genotype-environment interactions and, for gamete storage, a confounding of response and inbreeding effects are minimised.

Analysis of field records

In most farm animals, generations overlap so that field data can be utilised to provide estimates of genetic change unbiased by environmental fluctuation. In some early studies in dairy cattle, comparisons were made, within years, of age-corrected lactation records, but the estimates of age effects are confounded with any genetic or environmental trend (Rendel and Robertson, 1950). Recently, a method of estimating genetic change in broiler flocks, in which parts are replaced throughout the year, was proposed by Cassuto *et al.* (1970). Their analysis requires an estimate of the effect of maternal age on performance, which could similarly be confounded with any real trend.

In the methods commonly employed, the performance of contemporaneous progeny of sires born in different years is compared, a general technique suggested by Dickerson (1960). One method utilises least-squares analysis to estimate simultaneously the effects of bulls and years (Van Vleck and Henderson, 1961). In another, proposed by Smith (1962), the performance of the progeny of individual sires is compared with the mean of the whole population in each of several years. An estimate of annual genetic change in the population is obtained by doubling the regression of sire effects or contemporary comparison on years. The same principle is being used as in the long-term semen storage method described earlier, but the matings are no longer planned. For similar precision using field records, sires have to be used for many years and there has to be a considerable overlap of use of sires of different ages; usually large amounts of data are necessary. Other requirements and limitations of the method are similar to those noted for semen storage. In particular, data should not be included which are subsequently used to select sires on the basis of their progeny test, and there should be no association of mates' genotype or age with age of the sire. Lindström (1969) also summarised methods of predicting genetic changes from the selection differentials applied in the population, but such techniques must be distinguished from those which measure the response actually obtained.

Further discussion of ways of estimating change from field records are outside the scope of this review. The papers of Dickerson (1969) and Lindström (1969) should be consulted for more detail and references.

Efficiency of Methods of Estimating Change

Two types of error associated with any estimate of change must be distinguished, namely bias and sampling error. Most of the possible sources of bias associated with each method have been identified, but these are difficult to quantify, particularly in the case of genotype-environment interaction associated with some trend, or permanent change in the environment. In the review of McBride (1958) these are classed as " inter-population, macro-environmental " interactions, and in his review and the reviews of Falconer (1952, 1960), Dickerson (1962), Pirchner (1969), and Turner and Young (1969) the problems of genotype-environment interactions are discussed further. Presumably, interaction will be smaller when populations are more closely related or do not differ markedly for the traits under comparison, suggesting that a repeat mating design is preferable if such interactions are likely to be large. But under conditions of occasional major changes in the environment, a misleading result could be obtained with the repeat mating method if much selection is directed, in effect, towards adaptation to the current environment. In each generation, the selected line may appear better than the repeat mating, but after several generations could be no better than an unselected population. For example, imagine most selection is practised for resistance to a pathogen. If this mutates to a new form, all the gain made so far could be lost, but this would not be shown by the repeat mating control, which might record some generations of improvement followed by one of zero response.

Random sampling errors

The magnitude of random sampling errors can be established more exactly than that of bias (although difficulties are again encountered with genotype-environment interactions), so that the efficiencies of

alternative methods can be compared. Lasley (1960) and Dickerson (1969) discuss this problem, and other aspects have been considered by Hill (1971, 1972*a*, *b*) but partly from a different viewpoint. Dickerson was more concerned with estimating the actual genetic mean in the selected population, which would include genetic drift which had previously occurred in the selected line up to that time; Hill discussed the design of the whole selection experiment, specifically for realised heritability estimation, where the drift in the selected line is itself a source of error. There is little point in undertaking such a selection experiment using a very large control, with little drift or other source of variance, if the selected line itself is very small. The optimal allocation of resources will occur when roughly similar errors are found in both the selected line and the control. Further, where no common environmental changes are likely to occur, a control can be eliminated completely and all facilities devoted to selected lines. The following discussion is partly based on Dickerson's (1969) approach, but in a very simplified form; his excellent report should be consulted for more detail.

The possible sources of error in an estimate of change using a control population are: drift variance, σ^2_d, in the control, which increases roughly in proportion to generation number; error of measurement of the genetic mean from observations on phenotypes, σ^2_{ec} from the control and σ^2_{es} from the selected line, and genotype-environment interaction. The magnitude of the interaction between control and selected lines is difficult to predict *a priori*, so that the variance will be denoted σ^2_{it} at generation t. It is possible that the variance will not change as selection proceeds, in which case $\sigma^2_{it} = \sigma^2_{i0}$. Such might be the case with inbred lines, which comprise a small and special sample of genotypes. A more reasonable model is perhaps that in which the degree of interaction will increase in proportion to the genetic difference between individuals from the same or different populations; this is a consequence of regarding performance in different environments as correlated traits. If the response in a selected population is linear in t (generations), the variance between populations from genotype-environment interaction will increase in proportion to t^2 under this hypothesis, giving $\sigma^2_{it} = \sigma^2_{i0} + kt^2$, where k remains an unknown constant. Errors due to environmental trends or variation common to all individuals are removed by taking differences between the control and selected line means. Although the magnitude of alternative sources of variance may differ from scheme to scheme, the same symbols will be used for each design in order to demonstrate the error structure in a simple way.

Let $\bar{S}_t$ and $\bar{C}_t$ be the means of selected and control individuals at generation t. If the control and selected lines are drawn from a different base population, the response, to generation t, is estimated by $R_t = (\bar{S}_t - \bar{C}_t) - (\bar{S}_0 - \bar{C}_0)$, with variance V_C given by

$$V_C = 2\sigma^2_{es} + 2\sigma^2_{ec} + t\sigma^2_d + \sigma^2_{i0} + \sigma^2_{it}$$

If the lines are drawn from the same base population at the start of the experiment such that $\bar{S}_0 - \bar{C}_0 = 0$, the response can be estimated by $R_t = \bar{S}_t - \bar{C}_t$, with variance V'_C given by

$$V'_C = \sigma^2_{es} + \sigma^2_{ec} + t\sigma^2_d + \sigma^2_{it}$$

It is possible that the magnitude of the genotype-environment interaction term, σ^2_{it}, will be less in the second case, since the populations are more highly related. If genotypes (zygotes) can be stored, and used at intervals such that there is no drift variance associated with $\bar{C}_t$, the variance, V_Z, of R_t becomes

$$V_Z = 2\sigma^2_{es} + 2\sigma^2_{ec} + \sigma^2_{i0} + \sigma^2_{it}$$

or, if they are sampled from the same base population, V'_Z is given by V'_C without the term in σ^2_d. The same variance, V_Z, is associated with the use of completely inbred lines. If gametes are stored, an estimate of genetic change is obtained by doubling the observed response, so the variance of estimates of change, V_G, is given by $V_G = 4V_Z$. If the bulls from which semen is stored have accurately determined progeny tests, the appropriate error will be approximately $4V'_Z$.

In a repeat mating control of the type described by Goodwin *et al.* (1960), the responses each generation are estimated independently, so that if the control is used to estimate change in a different population, the measurement error variance, σ^2_{ec}, also accumulates in the control. The variance, V_R, of the estimate of genetic change in the selected population is now

$$V_R = 2\sigma^2_{es} + 2t\sigma^2_{ec} + t\sigma^2_d + \sigma^2_{i0} + \sigma^2_{it}.$$

If the repeat mating design is established within the population being selected, then the drift terms can be eliminated and the interaction variance should not increase. Interaction in intermediate generations is eliminated since each generation appears with a positive and negative sign in the estimate of response. Thus,

$$V'_R = t\sigma^2_{es} + t\sigma^2_{ec} + 2\sigma^2_{i0}$$

P. 6 L. 8 - 11. Replace by :

$$\sigma^2_{ec} = \frac{2\sigma^2}{N}\left[k + \frac{1 - h^2/2 - k}{n}\right]$$

and

$$\sigma^2_{ec}/\sigma^2_d = 4(nk + 1 - h^2/2 - k)/nh^2$$

Thus σ^2_{ec} exceeds σ^2_d if k is large or n and h^2 small, otherwise the two components will be .....

of individuals are recorded from the control and selected populations, and σ^2_{es} and σ^2_{ec} will be approximately equal. It is not possible to specify the order of magnitude of σ^2_t, the interaction which remains the real unknown in the calculations. The experimental evidence, to be reviewed in more detail in part II, indicates that the more highly related the populations, the smaller the interaction term, and theoretical arguments have been given above to suggest it may increase as populations diverge.

Estimation of change per generation

An estimate of genetic change per generation can be obtained in several ways. If the errors are uncorrelated with equal variance, the regression of cumulative response on generation number is most efficient, and is essentially equivalent to the regression of cumulative response on cumulative selection differential for realised heritability estimation. From Dickerson (1969) and Hill (1972*b*), the coefficients of the error terms in the regression using an unrelated control, *i.e.* taking terms from V_C, are as follows:

measurement (not accumulating)	$12(\sigma^2_{es} + \sigma^2_{ec})/[t(t+1)\ (t+2)]$,
drift (accumulating)	$1{\cdot}2(t^2+2t+2)\sigma^2_d/[t(t+1)\ (t+2)]$.

Therefore, as the duration of the experiment increases, the coefficients of σ^2_{ec} and σ^2_{es} in the regression of response on generation number decline in proportion to $1/t^3$, approximately, whereas that of σ^2_d declines in proportion to $1/t$. Thus, in a long-term experiment, most of the error from these two sources is contributed by drift. But with repeat mating controls, the measurement errors decline at the slow rate also. The effect of the interaction term is more equivocal. At best, if σ^2_{it} does not increase with t, its contribution will decline with $1/t^3$; at worst, if σ^2_{it} is proportional to t^2, the contribution of interaction to the variance of regression will *increase* in proportion to t. An alternative estimator of change per generation is simply the average response (*i.e.* R_t/t). Again, taking the terms from V_C, their coefficients are now:

measurement	$2(\sigma^2_{es} + \sigma^2_{ec})/t^2$,
drift	σ^2_d/t.

The average response is a slightly more efficient estimator than regression in the presence of drift variance, but poorer for measurement error, which declines in proportion to t^2. If σ^2_{it} increases in proportion to t^2, the interaction variance component in the average response is not dependent on t. Therefore, in the absence of interaction and unless the drift variance is very large, the regression estimator is superior. However, care must be taken in estimating the sampling variance of the regression of response on generations. Since the errors from drift are correlated and increase with successive generations, a biased estimate of variance is obtained from normal regression methods (Hill, 1972*a*, *b*).

Consider now alternative designs with roughly the same facilities devoted to each, firstly assuming interaction variance is unimportant, or non-increasing. Where possible, the genotype storage method is clearly most efficient if genotypes can be stored for the duration of the experiment, since there is no accumulating drift variance. Gamete storage has similar advantages. The control population is next most efficient, especially when steps are taken to minimise drift variance, and the repeat mating method, in which both the drift and measurement errors accumulate, is least efficient. If the repeat mating system is used in the selected line itself, there is no accumulation of drift, so it may be as efficient as a control population. However, one control population can be used for comparison with a number of selected lines, and thus requires many fewer facilities than repeat mating controls in each selected line.

Of course, the relative efficiency of the methods changes as the time span over which response is being evaluated alters. In a long-term experiment, drift, or other accumulating errors, contribute by far the greatest part of the variance; in one of only a few generations, the errors associated with estimating the mean may be more important. The problems of genotype-environment interaction and any possible trends in control performance are less easy to generalise; they are considered further in part II from experimental data. Where interaction variance is very large and increasing as populations diverge, the repeat mating method is likely to be most efficient. An alternative scheme is to use a control population, but to replace it every few generations, as the performance of the selected population increases, by a new control population with improved performance. Such a scheme has been proposed for pigs (Meat and Livestock Commission, 1970), and the error structure is now a mixture of that from control population and repeat mating systems.

Design of Control Populations

Several sources of error in estimation of genetic change using a control population (*i.e.* a segregating control population) have been identified: random genetic drift in the control, directional change in the control through natural or unintentional selection, interaction between the environment and the genotypes of the control and selected population and, finally, error of estimation of the control population mean through measuring few individuals. Error from the last source can be minimised if sufficient facilities are available, and it does not accumulate over generations. The other errors may accumulate, and so should be considered more carefully when establishing a population. These sources of error are discussed below. Particular attention is given to genetic drift because its magnitude can be quantified from *a priori* knowledge of the population structure, and no recent review directly relevant to control populations is available.

Random genetic drift

There is a vast literature on the theory of random genetic drift associated with finite population size, much of which has been summarised by Wright (1969) and Crow and Kimura (1970), but there have also been several more recent theoretical studies of drift in populations with overlapping generations. The theory is usually discussed in terms of the variance of gene frequencies, but here it is given for quantitative traits directly.

Consider a quantitative trait determined only by additive genes and having additive genetic variance σ^2_A. In an idealised random mating population of N monoecious (single sex) diploid individuals with discrete generations, and a random distribution of family sizes with no differential viability or fertility between families, the drift variance in a single generation is $\sigma^2_d = \sigma^2_A/N$. Over t generations the drift variance becomes $2\sigma^2_A\,[1-(1-1/2N)^t]$, which equals $t\sigma^2_A/N$, approximately, if N is large relative to t (*see e.g.* Crow, 1954; Falconer, 1960; Crow and Kimura, 1970). With non-additive gene action, these formulae in terms of σ^2_A no longer hold exactly, but are good approximations for dominant genes if t/N is small (Hill, 1972*a*). For populations with two sexes, other distributions of family size or non-random mating, the *effective population size* (or number), N_e, is defined as the number of individuals in the idealised population that would give the same drift in a single generation, *i.e.* $\sigma^2_d = \sigma^2_A/N_e$. The efficiency of alternative designs for control populations can thus be compared in terms of their effective size, a concept due to Wright (1931). Kimura and Crow (1963*a*) distinguish between a variance effective size, which is that defined above and predicts changes in drift variance, and an inbreeding effective size that predicts changes in heterozygosity. But in random mating populations of constant size, the two effective sizes are the same with non-overlapping generations (Kimura and Crow, 1963*a*) and with overlapping generations in models studied so far (Felsenstein, 1971; Crow and Kimura, 1971). Most control populations satisfy these restrictions and no distinction needs to be made between the alternative effective sizes; where non-random mating is discussed, only the variance effective size is considered.

To enable comparisons to be made between populations with different generation intervals, it is useful to define an *annual effective population size*, N_y. This is the size of an idealised population with a generation interval of one year which would give the same increment in drift variance *per year* as the population under consideration. In a population of generation interval L, in which the total drift in one year or generation is small, $N_y = LN_e$, and the increment in drift variance per year is $\sigma^2_A/N_y = \sigma^2_A/LN_e$. For example, a population of effective size 50 and generation interval of 2 years has the same annual effective size (100) as a population of effective size 100 and generation interval of 1 year. In species in which there are several generations per year a monthly effective size might be a more useful parameter.

Random mating and discrete generations. Formulae for effective size for cases of differing complexity, but with random mating and discrete generations, have been developed over many years (see Crow and Kimura, 1970). Some aspects can be clarified by using the monoecious model. The effective size is then

$$N_e^{-1} = (2+\sigma^2{}_n)/(4N-2) \qquad (1)$$

(Wright, 1938), where N is the actual size of the population and $\sigma^2{}_n$ the variance of family size. If there are no viability or fertility differences, such that family sizes vary only at random, then $\sigma^2{}_n = 2(1-1/N)$, a value differing slightly from that appropriate for the Poisson distribution, $\sigma^2{}_n = 2$, the average family size, because the total number of progeny over all families is constant. Thus with random sizes, $N_e = N$, and if all family sizes are equal, $\sigma^2{}_n = 0$ and $N_e = 2N-1$, from (1). In succeeding formulae, the population sizes will be assumed to be sufficiently large that equations such as (1) can be simplified, in this case to $N_e^{-1} = (2+\sigma^2{}_n)/4N$, and family sizes can be assumed to be Poisson distributed. Then, with random sizes, $\sigma^2{}_n = 2$ and $N_e = N$, as given by the exact formula, and with equal sizes, $\sigma^2{}_n = 0$ and $N_e = 2N$ rather than $2N-1$.

With two sexes, the effective size can be expressed in many ways. Based on a derivation of Latter (1959) for constant size and sex ratio, the formula is

$$\frac{1}{N_e} = \frac{1}{16M}\left[2+\sigma^2{}_{mm}+\frac{2M}{F}\,\mathrm{cov}(mm,mf)+\left(\frac{M}{F}\right)^2\sigma^2{}_{mf}\right] + \frac{1}{16F}\left[2+\left(\frac{F}{M}\right)^2\sigma^2{}_{fm}+\frac{2F}{M}\,\mathrm{cov}(fm,ff)+\sigma^2{}_{ff}\right] \qquad (2)$$

where M and F are the numbers of males and females. From male parents, the variance in the number of male progeny is $\sigma^2{}_{mm}$, of female progeny $\sigma^2{}_{mf}$, and the covariance of numbers of male and female progeny is $\mathrm{cov}(mm,mf)$; from female parents, the corresponding quantities are $\sigma^2{}_{fm}$, $\sigma^2{}_{ff}$ and $\mathrm{cov}(fm,ff)$. In the formulae of Kimura and Crow (1963*a*) and Crow and Kimura (1970), these covariance terms are omitted. This seems unjustified, for fertility differences between matings will usually produce a positive covariance in the number of male and female progeny. Latter (1959) considers a case where a negative covariance of family size is introduced for species such as poultry, in which a fixed total number of eggs may be set or progeny taken, from each family, regardless of sex.

Several well known examples illustrate the use of equation (2).

(*a*) With no viability or fertility differences between families, their sizes are Poisson distributed (approximately). Then $\sigma^2{}_{mm} = \sigma^2{}_{ff} = 1$, $\sigma^2{}_{mf} = 1/\sigma^2{}_{fm} = F/M$, $\mathrm{cov}(mm,mf) = \mathrm{cov}(fm,ff) = 0$, and $\frac{1}{N_e} = \frac{1}{4M} + \frac{1}{4F}$.

(*b*) As (*a*) but with an equal number ($N/2$) of males and females, then $N_e = N$.

(*c*) With equal numbers of males and females and every individual having one male and one female offspring, $N_e = 2N$. Thus, by removing one source of gene frequency drift, differential family sizes, the effective size is doubled, and the remaining drift comes solely from segregation within heterozygotes.

(*d*) In the control population design described by King *et al.* (1959) and Gowe *et al.* (1959), each male has one son and F/M daughters, and each female has one daughter and a probability of M/F of having one son. Then $\sigma^2{}_{fm} = \frac{M}{F}\left(1-\frac{M}{F}\right)$, other variances and covariances of family number are zero, and

$$\frac{1}{N_e} = \frac{3}{16M} + \frac{1}{16F}$$

With 50 males and 250 females the effective size is now 250, compared with 167 when family sizes are random (as (*a*) above). Thus, steps should always be taken in pedigreed populations to equalise family sizes.

Random mating and overlapping generations. Various formulae for effective population sizes with overlapping generations have been derived in the past few years: some (Kimura and Crow, 1963*a*; Nei and Imaizumi, 1966; Giesel, 1969) have later been considered incorrect or vague (see Felsenstein, 1971; Hill, 1972*c*) and others relate to special models, either haploid (Moran, 1962) or with very large numbers of females (Turner and Young, 1969). The problems have been clarified in recent studies by Felsenstein (1969, 1971), Nei (1970), Crow and Kimura (1971) and Hill (1972*c*). In a control population which is properly managed, the number of animals entering the herd each year and the age distribution of individuals in the herd should remain constant, or at least show little variation. Similarly, the age

distribution of parents of individuals born in any year should also be stable. With these conditions, the effective size of a population with overlapping generations is the same as that of a population with discrete generations, and is given by equation (2), in which M and F are the number of males and females entering the population each *generation* and the variances and covariances, $\sigma^2_{m\,m}$, cov(mm,mf), etc., are of *lifetime* family size (Hill, 1972*c*). The generation interval here is defined in the usual way, as the average age of parents when progeny are born in the four pathways male to male, etc. (Rendel and Robertson, 1950). In any newly established population, the increment in drift variance will not be exactly σ^2_A/N_e until the age distribution and relationships of individuals in the population have stabilised.

If family sizes are Poisson distributed, $N_e=4MF/(M+F)$ as in the discrete model, and the annual effective size is $N_y=4MFL/(M+F)$. In terms of the numbers of males and females entering per year, m and f, say, $N_y=4mfL^2/(m+f)$, and there are obvious advantages in increasing generation interval. Of course, if the number entering per year is fixed, an increase in generation interval implies an increase in the total size of the population, a parameter which does not appear in this formulation. However, if females have to be maintained a long time, merely to replace the population when the birth rate is low, or if accommodation is limiting only for females, it may be worth while to replace males rapidly (Turner and Young, 1969). For example, imagine a control flock of sheep in which the mean age of females when their progeny are born is 3 years. The flock is to be maintained with K males, many less than the number of females, and the males can have their first progeny when 1 year of age. With average generation interval L and m males entering per year, the annual effective size is $N_y=4mL^2$ if the number of progeny per male is Poisson distributed. Consider 3 strategies:

(a) Males used once: $m=K$, $L=2$, $N_y=16K$
(b) Males used twice: $m=K/2$, $L=2{\cdot}25$, $N_y=10{\cdot}1K$
(c) Males used 3 times: $m=K/3$, $L=2{\cdot}5$, $N_y=8{\cdot}3K$.

These annual effective sizes could be doubled by ensuring each male was replaced by one son. Other strategies, for example using males for 1 year only, but such that their progeny are born when they are 2 years of age, require the maintenance of twice as many adult or young males in the flock. Other systems in which females are kept longer may increase the generation interval, but at the same time could increase the variance of family size through differential survival.

Perhaps since control populations have been used primarily in poultry and laboratory animals, which typically are reproduced with discrete generations, there seems to be little discussion in the literature on design with overlapping generations, other than the brief study by Turner and Young (1969). Further analyses of possible structures for control populations with overlapping generations are required. Some assumptions normally made with discrete generations are less tenable: if there is a Poisson distribution of family size among surviving breeding individuals which is compounded with differential survival, even without selection, the lifetime distribution of family size is no longer Poisson. For example, if age at death follows an exponential distribution, the variance of lifetime family size is three times its mean and the effective population size one half the value it would be if lifetime family sizes were Poisson distributed (Moran, 1962; Felsenstein, 1971; Hill, 1972*c*). Therefore steps should be taken when maintaining a control to reduce differential survival of breeding individuals, and, as with discrete generations, attempt to equalise lifetime family size.

Non-random mating. Genetic drift may be reduced by practising non-random mating of individuals on the basis of their relationship to each other (Kimura and Crow, 1963*b*; Robertson, 1964; Wright, 1965; Cockerham, 1967, 1970). If family sizes are equal, the increment in variance of gene frequency is proportional to the heterozygosity in the previous generation (Kimura and Crow, 1963*b*). Therefore, any mating system which minimises the amount of heterozygosity also minimises the gene frequency drift. In practice, this requires that mates should be more closely related than the average relationship within the population (Robertson, 1964), so for minimum drift the population should be broken up as rapidly as possible into the maximum number of sublines, which will contain highly inbred animals, and these sublines should be permanent. However, any degree of sublining will reduce the increase in genetic drift. Further, as with random mating populations, equal numbers of progeny should be raised in each family in order to minimise the drift, and this restriction is assumed to be practised in the following systems.

In breeding schemes in which there is " maximum avoidance " of mating relatives there is an increase in heterozygosity over random mating schemes, and consequently an increase in drift. The variance effective size is reduced by about $\frac{1}{2}\log_2 N$ below $2N$, the value for random mating, where N is an integral power of 2 (Robertson, 1964), so the effect is rather small. If only full-sib matings are avoided, Robinson and Bray (1965) found the effective population size to be reduced by 1 if family sizes are equal, whereas

Jacquard (1971) concluded there was no reduction, but when family sizes are randomly distributed, the effective size is increased by $1\frac{1}{2}$ above N (Robinson and Bray, 1965; Jacquard, 1971).

Kimura and Crow (1963*b*) describe a circular mating design in which the males and females are conceptually arranged on a circle, and each individual mated with both its neighbours. The initial increase in drift is not regular, but eventually asymptotes at a value corresponding to an effective size of $2(N+2)^2/\pi^2$, approximately. If only pair matings are made, the effective size is $(N+12)^2/2\pi^2$. These and other circular mating schemes considered by Kimura and Crow (1963*b*) and Maruyama (1970), in which the effective population size is proportional to the square of the actual size, can be regarded as a special type of sublining in which the mating system follows the same plan each generation (Robertson, 1964). For example, with the first type of circular mating described, each individual is mated to a half-sib. But over a period of a few generations, the drift variances accumulated in the alternative schemes (random mating, maximum avoidance circular mating and sublining) will differ very little if the population size is not too small. The most important restriction is to keep family sizes equal.

All of the above schemes involving non-random matings are of questionable value for control populations. In each case, an increase in effective size is obtained by mating relatives and consequently increasing the inbreeding level of breeding individuals. There might then be difficulties in reproducing the line without unconscious selection, and for testing purposes crosses would need to be made between, say, sublines, and extra facilities would be required.

Finally, two alternative methods of influencing the drift variance are considered in more detail, for they are not given in the control population context elsewhere.

Negative assortative mating. If negative assortative mating is practised on performance of some trait, its genetic variance is expected to be reduced (Fisher, 1918; Wright, 1921; Crow and Felsenstein, 1968; Crow and Kimura, 1970). This could be useful in a control population, since the drift variance for this particular trait should be reduced in proportion, but would be of little value if the control were being used as a standard for several traits.

With a phenotypic correlation of r among mates imposed for a trait of heritability h^2, the additive genetic correlation among mates is rh^2 and the additive variance of the trait becomes $(1+rh^2/2)\sigma^2_A$ after one generation. With perfect negative assortative mating ($r=-1$) the variances would be $0{\cdot}75\sigma^2_A$ for $h^2=0{\cdot}5$ and $0{\cdot}875\sigma^2_A$ for $h^2=0{\cdot}25$. With continued assortative mating, some further reductions occur in later generations. Crow and Kimura (1970) show that if the trait is additive and affected by a large number of independent loci, the additive genetic correlation among mates, a, is given by the solution to

$$(1-h^2)a^2-a+h^2r=0$$

and the additive variance stabilised at $\sigma^2_A/(1-a)$, where h^2 and σ^2_A describe the population prior to the assortative mating. With $r=-1$, the additive variance asymptotes at $0{\cdot}71\sigma^2_A$ for $h^2=0{\cdot}5$ and $0{\cdot}82\sigma^2_A$ for $h^2=0{\cdot}25$, so most of the reduction in variance occurs immediately. The net result is to increase the effective size, relative to a random mating population, by about 20% with no increase in facilities required. With tightly linked loci there is a further reduction in the final variance, but it takes longer to achieve; while with fewer loci of larger effects, the reduction in variance is expected to be smaller.

The assortative mating acts by reducing the variance between families. This reduction can also be achieved by choosing individuals with as small a range of phenotypes as possible; these should have performance near the mean if no directional change is to be produced (see Bulmer, 1971, for an analysis). The practical efficiency of these methods is questionable: negligible effects on variance were observed experimentally by Falconer (1957).

Zero selection differential. If the performance of more individuals is measured than are used for reproducing the population, the breeding individuals can be chosen such that their mean performance for some particular trait is close to the mean performance of all recorded individuals in that generation. By ensuring a selection differential of zero, or nearly zero, the drift variance for this trait can be reduced; the technique has been used in practice (Turner *et al.*, 1968; Edwards *et al.*, 1971). It is not essential that individuals with performance near the mean be chosen; if so, or if negative assortative mating is practised among the selected individuals, a further reduction in variance should result. The necessary theory is given by Hill (1971), but in the context of directional selection.

When a group of N individuals is taken at random from a random mating population, the variance of their mean genotype, or drift variance of their progeny mean, is σ^2_A/N. Conditional on the deviation of the observed mean of these N individuals from the population mean, the variance is reduced to $(1-h^2)\sigma^2_A/N$, for this is now a variance about regression where the heritability, h^2, is the square of the correlation of genotype and phenotype. If only a finite number of individuals is measured, such that

the N selected comprise a fraction p of them, the true population mean is no longer known exactly and the variance increases to $[1-h^2(1-p)]\sigma^2_A/N$.

Consider now a control population in which family sizes have a Poisson distribution, and breeding individuals are chosen such that the selection differential is zero but with roughly a Poisson distribution of number from each family. The drift variance is then

$$\sigma^2_d=\frac{\sigma^2_A}{4}\left\{\frac{1}{M}[1-h^2(1-p_m)]+\frac{1}{F}[1-h^2(1-p_f)]\right\}$$

where p_m and p_f are the proportions of males and females chosen from those recorded. For $p_m=p_f=p$, the formula simplifies, and the effective population size is given by

$$\frac{1}{N_e}=\left(\frac{1}{4M}+\frac{1}{4F}\right)[1-h^2(1-p)]$$

If a very large number is recorded ($p\to 0$), the effective size is increased by $1/(1-h^2)$, and is doubled if $h^2=0{\cdot}5$. When selection is practised on some index, such as deviation from family average, having correlation r_{IG} with breeding value, the appropriate effective size is multiplied by a factor $1/[1-r^2_{IG}(1-p)]$ (from Hill, 1971). If there are $N/2$ full-sib families and equal family sizes are used, the effective size is $2N$ without selection, and with selection within families r_{IG} depends on family size. If families are large (and consequently p is small), $r^2_{IG}=h^2/4(1-\frac{1}{2}h^2)$ and the effective size is increased by a factor of only 1·043 if $h^2=0{\cdot}5$. Finally, when selection is practised within half-sib families to ensure that each sire has only one son, $r^2_{IG}=9h^2/4(4-h^2)$ if family sizes are large, and the effective size is multiplied by 1·32 if $h^2=0{\cdot}5$. It should be possible in practice to choose individuals such that the selection differential is close to zero for several traits, particularly when many individuals are recorded. The technique therefore appears worth while if family sizes are random, or when restrictions are placed only on the replacement of sires by sons. If full-sib families are represented equally among the parents there is little further advantage in ensuring a zero selection differential.

Directional genetic change

In segregating control populations a directional trend in performance can arise from several causes: inbreeding depression, natural selection or, temporarily, epistatic loci initially in linkage disequilibrium. Since the variance and inbreeding effective numbers are the same in unselected random mating populations (see above), the inbreeding depression is also minimised in control populations designed to minimise genetic drift, providing they are random mating. However, some of the schemes which utilise non-random mating may quickly give high levels of inbreeding, and are therefore not applicable if the control population is to be used for traits showing much inbreeding depression. Alternatively, crosses between replicated controls can be made whenever comparisons are required, so that there is then no inbreeding. For example, if the available facilities are devoted to two populations, but the total inbreeding in each remains low, the genetic drift in each will be nearly double that in a single population of twice the actual size maintained with the same breeding programme. The drift in the mean performance of the two will be almost as high as in the single large population, but their cross will show no inbreeding depression. Of course, more facilities are required for rearing the crosses. If many small sublines are maintained, the drift will be appreciably lower in their final cross than in a single large population (see above). Alternatively, the control population can be maintained with the same effective size as the selected population, so that about the same amount of inbreeding depression will occur in both. But it will be difficult to ensure the effective sizes are the same, since selection may reduce the drift variance (Hill, 1971; or see section on zero selection differential) but can also increase it through an increase in the variance of family size (Robertson, 1961). Therefore the effective sizes may be most nearly equal when within-family selection is practised. Further, the selection itself will change gene frequencies and thus may alter the subsequent inbreeding depression of a quantitative trait per unit inbreeding.

The effects of natural selection acting directly or through pleiotropy on the traits of interest are less predictable, and although steps can be taken to minimise these effects, it may be impossible or unduly laborious to eliminate them entirely (Gowe *et al.*, 1959). Ideally there should be no differential fertility or mortality, or at least no correlation of these traits with any trait of interest. Gowe *et al.* noted that by ensuring equality of family size, there is no selection based on fitness differences between families. This may be easier in species such as poultry, where there is spare reproductive capacity, than in others

such as sheep. If there are some completely infertile animals, Gowe *et al.* suggest that these be replaced by their full-sibs wherever possible. Since the effective size of the population is increased by ensuring that family sizes are equal, the aims of minimising between-family selection and drift variance are completely compatible. Selection within families as a result of differential mortality cannot be entirely avoided, especially at the early embryo stage, but steps should be taken to ensure survival of as high a proportion of individuals as possible (Gowe *et al.*, 1959).

A control population formed by relaxation of a selected population may undergo initial regression if there are epistatic loci affecting the trait which are not in linkage equilibrium. Unless linkage is tight almost half the additive × additive effects (Griffing, 1960) and three-quarters or more of higher order interactions will be lost in the first generation. The problem has been discussed by Dickerson (1965), who also provides evidence of this regression (or recombination loss) in poultry. Natural selection, if important, is also likely to have more effect in recently relaxed populations than in those which have not been under selection for some time. For example, genes have been found in experiments on *Drosophila* which reach high frequency in selected lines, but have highly deleterious effects on fitness and are rapidly reduced in frequency on relaxation (Clayton and Robertson, 1957). Sometimes a comparison of a selected line and a relaxed line drawn from the same commercial population is required, for this difference measures the returns from the selection practised in that particular population and should include any regression on relaxation.

Conclusions

It is clear that if steps are taken to keep family sizes equal, both drift variance and possible directional selection effects are minimised. But the magnitude of effects of natural selection and of genotype-environment interactions are difficult to quantify from theoretical arguments. Estimates of their real importance in practical situations can only be obtained from experimental analysis of field data. One of the main objectives of part II of this study is to review this information on control population stability.

Acknowledgements

The author is indebted to Prof. A. Robertson, Dr. R. B. Land and Dr. Y. Yamada for their valuable comments and suggestions.

REFERENCES

A.B.A. = Animal Breeding Abstracts

BELL, A. E., MOORE, C. H., and WARREN, D. C. 1955. The evaluation of methods for the improvement of quantitative characteristics. *Cold Spring Harb. Symp. quant. Biol.*, **20**: 197–212. [*A.B.A.*, **25,** No. 1545.]

BRAY, D. F., BELL, A. E., and KING, S. C. 1962. The importance of genotype by environment interaction with reference to control populations. *Genet. Res.*, **3**: 282–302. [*A.B.A.*, **30,** No. 2913.]

BULMER, M. G. 1971. The effect of selection on genetic variability. *Am. Nat.*, **105**: 201–211. [*A.B.A.*, **40,** No. 1145.]

CASSUTO, D., BELL, A. E., and ANDERSON, V. L. 1970. Estimation of genetic gains in populations with overlapping generations. *Br. Poult. Sci.*, **11**: 217–230. [*A.B.A.*, **38,** No. 4286.]

CLAYTON, G. A. 1968. Some implications of selection results in poultry. *Wld's Poult. Sci. J.*, **24**: 37–57. [*A.B.A.*, **38,** No. 4287.]

CLAYTON, G. A., MORRIS, J. A., and ROBERTSON, A. 1957. An experimental check on quantitative genetical theory. I. Short-term responses to selection. *J. Genet.*, **55**: 131–151. [*A.B.A.*, **25,** No. 1550.]

CLAYTON, G. A., and ROBERTSON, A. 1957. An experimental check on quantitative genetical theory. II. The long-term effects of selection. *J. Genet.*, **55**: 152–170. [*A.B.A.*, **25,** No. 1550.]

COCKERHAM, C. C. 1967. Group inbreeding and coancestry. *Genetics, Austin, Tex.*, **56**: 89–104. [*A.B.A.*, **35,** No. 4161.]

COCKERHAM, C. C. 1970. Avoidance and rate of inbreeding. In *Mathematical topics in population genetics. Ed. by K. Kojima. Biomathematics*, **1**. *Berlin, Heidelberg and New York: Springer-Verlag.* Pp. 104–127.

CROW, J. F. 1954. Breeding structure of populations. II. Effective population number. In *Statistics and mathematics in biology. Ed. by O. Kempthorne et al. Ames: Iowa State College Press.* Pp. 543–556.

CROW, J. F., and FELSENSTEIN, J. 1968. The effect of assortative mating on the genetic composition of a population. *Eugen. Q.*, **15**: 85–97. [*A.B.A.*, **37,** No. 3059.]

CROW, J. F., and KIMURA, M. 1970. An introduction to population genetics theory. *New York, Evanston and London: Harper & Row, Publishers.* xiv + 591 pp.

CROW, J. F., and KIMURA, M. 1971. The effective number of a population with overlapping generations: A correction and further discussion. *Am. J. hum. Genet.* (In press.)

DICKERSON, G. E. [? 1960.] Techniques for research in quantitative animal genetics. In *Techniques and procedures in animal production research. Beltsville, Md: American Society of Animal Production.* Pp. 56–105.

DICKERSON, G. [E.] 1961. Effectiveness of selection for animal improvement. In *Germ plasm resources. Publs Am. Ass. Advmt Sci.*, No. 66: 161–190. [*A.B.A.*, **30,** No. 1338.]

DICKERSON, G. E. 1962. Implications of genetic-environmental interaction in animal breeding. *Anim. Prod.*, **4**: 47–64. [*A.B.A.*, **30,** No. 1544.]

DICKERSON, G. E. 1965. Experimental evaluation of selection theory in poultry. In *Genetics today. Proc. XIth int. Congr. Genet., The Hague*, 1963, Vol. 3: 747–760.

DICKERSON, G. E. [? 1968.] Lessons to be learned from poultry breeding. In *Animal breeding in the age of AI. Symp. co-sponsored by Univ. Wisconsin Coll. Agric. Life Sci. and Am. Breed. Serv., Inc., Feb. 29–Mar. 1, 1968, Madison, Wis. Madison.* Pp. 69–93, 94–99.

DICKERSON, G. E. 1969. Techniques for research in quantitative animal genetics. In *Techniques and procedures in animal science research.* Revised edn. *New York: American Society of Animal Science.* Pp. 36–79.

EDWARDS, R. L., OMTVEDT, I. T., and WHATLEY, J. A. 1971. Genetic analysis of a swine control population. I. Population stability. *J. Anim. Sci.*, **32**: 179–184. [*A.B.A.*, **39,** No. 3617.]

FALCONER, D. S. 1952. The problem of environment and selection. *Am. Nat.*, **86**: 292–298. [*A.B.A.*, **23,** No. 940.]

FALCONER, D. S. 1957. Selection for phenotypic intermediates in *Drosophila*. *J. Genet.*, **55**: 551–561.

FALCONER, D. S. 1960. Introduction to quantitative genetics. *Edinburgh and London: Oliver & Boyd Ltd.* ix + 365 pp.

FALCONER, D. S., and LATYSZEWSKI, M. 1952. The environment in relation to selection for size in mice. *J. Genet.*, **51**: 67–80. [*A.B.A.*, **21,** No. 332.]

FELSENSTEIN, J. 1969. The effective size of a population with overlapping generations. Abstr. in *Genetics, Austin, Tex.*, **61** (No. 2: Pt 2, Suppl., *A. Rep. Issue Genet. Soc. Am.*): s18.

FELSENSTEIN, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics*. (In press.)

FISHER, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, **52**: 399–433.

GIESBRECHT, F., and KEMPTHORNE, O. 1965. Examination of a repeat mating design for estimating environmental and genetic trends. *Biometrics*, **21**: 63–85. [*A.B.A.*, **34,** No. 808.]

GIESEL, J. T. 1969. Inbreeding in a stationary, stable population as a function of age and fecundity distribution. Abstr. in *Genetics, Austin, Tex.*, **61** (No. 2: Pt 2, Suppl., *A. Rep. Issue Genet. Soc. Am.*): s21.

GOODWIN, K., DICKERSON, G. E., and LAMOREUX, W. F. 1955. A technique for measuring genetic progress in poultry breeding experiments. Abstr. in *Poult. Sci.*, **34**: 1197.

GOODWIN, K., DICKERSON, G. E., and LAMOREUX, W. F. 1960. An experimental design for separating genetic and environmental changes in animal populations under selection. In *Biometrical genetics. Ed. by O. Kempthorne. Un. int. Sci. biol. Ser. B (Colloq.)*, No. 38: 117–138.

GOWE, R. S., ROBERTSON, A.. and LATTER, B. D. H. 1959. Environment and poultry breeding problems. 5. The design of poultry control strains. *Poult. Sci.*, **38**: 462–471. [*A.B.A.*, **28,** No. 362.]

GRIFFING, B. 1960. Theoretical consequences of truncation selection based on the individual phenotype. *Aust. J. biol. Sci.*, **13**: 307–343.

HICKMAN, C. G. 1958. Population dynamics in dairy cattle and a measure of genetic change. *Proc. Genet. Soc. Can.*, **3** (2): 3–6. [*A.B.A.*, **28,** No. 600.]

HICKMAN, C. G., and FREEMAN, A. E. 1969. New approach to experimental designs for selection studies in dairy cattle and other species. *J. Dairy Sci.*, **52**: 1044–1054. [*A.B.A.*, **38,** No. 249.]

HILL, W. G. 1971. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics*, **27**: 293–311. [*A.B.A.*, **40,** No. 1154.]

HILL, W. G. 1972*a*. Estimation of realised heritabilities from selection experiments. I. Divergent selection. *Biometrics*, **28**. (In press.)

HILL, W. G. 1972*b*. Estimation of realised heritabilities from selection experiments. II. Selection in one direction. *Biometrics*, **28**. (In press.)

HILL, W. G. 1972*c*. Effective size of populations with overlapping generations. *Theor. Popul. Biol.* (Submitted.)

JACQUARD, A. 1971. Effect of exclusion of sib-mating on genetic drift. *Theor. Popul. Biol.*, **2**: 91–99.

JAMES, E. 1961. Perpetuation and protection of germ plasm as seed. In *Germ plasm resources. Publs Am. Ass. Advmt Sci.*, No. 66: 317–326.

KIMURA, M., and CROW, J. F. 1963*a*. The measurement of effective population number. *Evolution, Lawrence, Kans.*, **17**: 279–288. [*A.B.A.*, **32,** No. 1613.]

KIMURA, M., and CROW, J. F. 1963*b*. On the maximum avoidance of inbreeding. *Genet. Res.*, **4**: 339–415. [*A.B.A.*, **32,** No. 1614.]

KING, S. C., CARSON, J. R., and DOOLITTLE, D. P. 1959. The Connecticut and Cornell randombred population of chickens. *Wld's Poult. Sci. J.*, **15**: 139–159. [*A.B.A.*, **28,** No. 989.]

KOJIMA, K., and KELLEHER, T. M. 1963. A comparison of purebred and crossbred selection schemes with two populations of *Drosophila pseudoobscura*. *Genetics*, **48**: 57–72.

LARSON, R. E. 1961. Perpetuation and protection of germ plasm as vegetative stock. In *Germ plasm resources. Publs Am. Ass. Advmt Sci.*, No. 66: 327–337.

LASLEY, E. L. 1960. Lessons from experiments with control populations of laboratory organisms. *Proc. 9th U.S. Natn. Poult. Breeders Round Table.*

LATTER, B. D. H. 1959. Genetic sampling in a random mating population of constant size and sex ratio. *Aust. J. biol. Sci.*, **12**: 500–505. [*A.B.A.*, **28,** No. 1673.]

LERNER, I. M. 1950. Population genetics and animal improvement. *Cambridge: University Press.* xviii + 342 pp.

LINDSTRÖM, U. 1969. Genetic change in milk yield and fat percentage in artificially bred populations of Finnish dairy cattle. *Suom. maatal. Seur. Julk.*, **114**: 128 pp. [*A.B.A.*, **38,** No. 2381.]

MCBRIDE, G. 1958. The environment and animal breeding problems. *Anim. Breed. Abstr.*, **26**: 349–358.

MARUYAMA, T. 1970. Rate of decrease of genetic variability in a subdivided population. *Biometrika*, **57**: 299–311. [*A.B.A.*, **39,** No. 1270.]

MEAT AND LIVESTOCK COMMISSION. 1970. Pig improvement. Report of a Scientific Study Group. *London: Meat and Livestock Commission.*

MORAN, P. A. P. 1962. The statistical processes of evolutionary theory. *Oxford: At the Clarendon Press. London: Oxford University Press.* vii+200 pp.

NEI, M. 1970. Effective size of human populations. *Am. J. hum. Genet.*, **22**: 694–695.

NEI, M., and IMAIZUMI, Y. 1966. Genetics structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. *Heredity, Lond.*, **21**: 183–190, 344.

PIRCHNER, F. 1969. Population genetics in animal breeding. *San Francisco, Calif.: W. H. Freeman & Co.* xi+274 pp.

RAHNEFELD, G. W., BOYLAN, W. J., COMSTOCK, R. E., and SINGH, M. 1963. Mass selection for post-weaning growth in mice. *Genetics, Austin, Tex.*, **48**: 1567–1583. [*A.B.A.*, **32,** No. 1357.]

RENDEL, J. M., and ROBERTSON, A. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. *J. Genet.*, **50**: 1–8. [*A.B.A.*, **18,** No. 943.]

ROBERTS, R. C. 1965. Some contributions of the laboratory mouse to animal breeding research. Pt I. Pt II. *Anim. Breed. Abstr.*, **33**: 339–353; 515–526.

ROBERTSON, A. 1961. Inbreeding in artificial selection programmes. *Genet. Res.*, **2**: 189–194. [*A.B.A.*, **30,** No. 1419.]

ROBERTSON, A. 1964. The effect of non-random mating within inbred lines on the rate of inbreeding. *Genet. Res.*, **5**: 164–167. [*A.B.A.*, **32,** No. 2569.]

ROBINSON, P., and BRAY, D. F. 1965. Expected effects on the inbreeding coefficient and rate of gene loss of four methods of reproducing finite diploid populations. *Biometrics*, **21**: 447–458.

SITTMANN, K., ABPLANALP, H., and FRASER, R. A. 1966. Inbreeding depression in Japanese quail. *Genetics, Austin, Tex.*, **54**: 371–379. [*A.B.A.*, **35,** No. 2013.]

SMITH, C. 1962. Estimation of genetic change in farm livestock using field records. *Anim. Prod.*, **4**: 239–251. [*A.B.A.*, **30,** No. 2322.]

TURNER, Helen Newton, DOLLING, C. H. S., and KENNEDY, J. F. 1968. Response to selection in Australian Merino sheep. I. Selection for high clean wool weight, with a ceiling on fibre diameter and degree of skin wrinkle. Response in wool and body characteristics. *Aust. J. agric. Res.*, **19**: 79–112. [*A.B.A.*, **36,** No. 2700.]

TURNER, Helen Newton, and YOUNG, S.[S. Y.] 1969. Quantitative genetics in sheep breeding. *South Melbourne, Vict.: Macmillan Co. of Asutralia Pty Ltd. London: Macmillan & Co. Ltd. Ithaca, N.Y.: Cornell University Press.* xviii+332 pp.

VAN VLECK, L. D., and HENDERSON, C. R. 1961. Measurement of genetic trend. *J. Dairy Sci.*, **44**: 1705–1710. [*A.B.A.*, **30,** No. 962.]

WRIGHT, S. 1921. Systems of mating. *Genetics, Princeton,* **6**: 111–178.

WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics, Princeton,* **16**: 97–159.

WRIGHT, S. 1938. Size of population and breeding structure in relation to evolution. *Science, N.Y.*, **87**: 430–431.

WRIGHT, S. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution, Lawrence, Kans.*, **19**: 395–420. [*A.B.A.*, **34,** No. 2657.]

WRIGHT, S. 1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies. *Chicago and London: University of Chicago Press.* vii+511 pp.

16

Investment appraisal for national breeding programmes

by

William G. Hill

Anim. Prod. 1971, 13: 37–50

INVESTMENT APPRAISAL FOR NATIONAL BREEDING PROGRAMMES

W. G. HILL

Institute of Animal Genetics, Edinburgh EH9 3JN

SUMMARY

The discounted cash flow procedure of management accounting is used to evaluate national breeding programmes. Alternative methods of improvement of meat production from cattle born in the dairy herd are taken as examples. These schemes utilize selection for beef characters either in the dairy breed itself or within a beef breed, maintained in a small herd and used as a source of bulls for crossing by AI. In each case young bulls are selected for growth rate in a performance test, which precedes the progeny test for milk production in the dairy breed.

Greater rates of genetic progress and monetary returns are predicted from improvement of the beef breed, but both schemes are expected to yield a return on investment of over 15%. The net returns from the programme in the beef breed are influenced less by changes in assumptions.

Approximate, but simple, methods of computing the discounted returns are described.

INTRODUCTION

THE efficiency of any breeding programme can be judged in many ways. The geneticist may be concerned solely with the rates of improvement which can be achieved and these can be predicted if there is sufficient information on parameters such as heritabilities. But alternative programmes involve different costs, both in the initial capital required for testing facilities and stock and in annual running expenditure for recording, maintaining large numbers of entire males, carcass dissection and so on. The monetary returns from genetic improvement accumulate over a long period of time, and the pattern of returns may be erratic in early years while the genes from selected animals are distributed through the population. There is clearly a need for some rational method of combining the returns and costs for any programme so that sound investment decisions can be made. The economic objectives may differ among breeders. Commercial companies must measure their success largely in terms of the proportion of the market for breeding stock which they are able to capture, and since there is unlikely to be a linear relationship between the merit of their stock and their share of the market or profitability, investment policy is difficult to analyse. For national breeding organizations, such as the Meat and Livestock Commission in Great Britain, the competitive element may be less important and the problem of evaluating returns from the selection programme is less equivocal. Nevertheless, providing importation of stock from abroad is permitted, competition exists among national organizations for the use of their

breeding animals. Further, the genetic merit of animals within any country affects the efficiency of agricultural production and hence the size of the agricultural industry and market for breeding stock. In the analysis to follow these problems will be ignored and a relatively stable size of the production industry assumed. However, even a national organization has to compete for finance, so more is required than just a scheme for comparing breeding programmes. Overall returns on investment must also be predicted so that they can be considered along with extreme alternatives such as the building of a new road. There are, of course, social factors which also govern expenditure in particular areas; these too will be ignored.

Use will be made of a standard technique in management accounting, the discounted cash flow method. This discounting procedure has already been used in studies of selection programmes for dairy cattle and dual purpose dairy-beef populations (Poutous and Vissac, 1962; Soller, Bar-Anan and Pasternak, 1966; Lindhé, 1969; Hinks, 1970a,b), while the examples in this paper will refer solely to improvement of meat production in cattle. Emphasis will be given to the accounting and decision-making procedure and its implications rather than to recommendations in particular situations, and the genetic content of the paper is small. However, the time scale of obtaining genetic improvement in the population in any scheme is irregular, as Searle (1961) has shown for dairy cattle improvement. This irregularity is highlighted in the discounting method.

By necessity, a large number of simplifying assumptions are made.

DISCOUNTING PROCEDURE

The discounted cash flow technique is discussed fully in texts on management accounting and a relevant summary is given by the British Treasury in a memorandum (House of Commons, 1967–8) concerning investment analysis for Nationalized Industries. Some of the arguments will now be summarized.

Let us assume that the current rate of interest is 8%. Then £100 invested now at compound interest would become £108 next year, £100 $\times (1{\cdot}08)^2$ the following year and so on. Conversely, £108 obtained next year is equivalent to receiving £100 now, or £1 next year is worth £1/1·08 = £0·9259 now, and £1 in two years time is equivalent to £$(0{\cdot}9259)^2$ = £0·8573 this year. With an 8% discount rate one monetary unit obtained at years 5, 10 and 20 represents 0·6806, 0·4632 and 0·2145 units at current value. In this way all expenditure and returns made in different years can be equated to the same base year, and by summation an aggregate profit computed up to any year. Thus breeding programmes which lead to very different time patterns of returns can be compared in a simple way.

All expenses, including those regarded as capital, are included in the year they are incurred. In practice a breeding programme might be continued indefinitely; however, we need to specify a period of, say, 15 or 20 years over which the investment is to be judged. At the end of this period all realizable assets, such as land or stock in this context, are 'sold off' and counted as returns in that year. No provision for depreciation is then necessary. Inflation of monetary value is not included, for it is assumed that the rate is similar for costs and returns in the breeding programme and in the economy as a whole.

ANALYSIS OF CATTLE IMPROVEMENT SCHEMES

Models

We shall consider two schemes for improvement of meat production from cattle in which the dairy herd is the source of calves. The dairy breed is assumed to be of dual-purpose type, and purebred animals, as well as crosses with beef breeds are reared and marketed at slaughter weights around 400–600 kg. Beef production can therefore be increased both by selection within the dairy breed and by selection of the beef breed or breeds used for crossing with it. All matings in the dairy breed are assumed to be by AI, and a progeny testing scheme for milk production is already in operation. There are two main alternatives for improving traits of meat production in the dairy breed: a performance test of young bulls prior to the milk progeny test, or including beef characters in the progeny test itself. These alternatives were compared by Soller *et al.* (1966) and since they found the performance testing scheme more efficient we shall consider it alone. Rather few beef bulls are needed for crossing by AI, so these can be supplied from a single closed herd of one breed. In this herd a performance testing scheme is operated with a short generation interval. Selection on males alone is practised, since roughly the same rates of progress can be achieved as with selection on females also (D. E. Steane, personal communication). For example, with the wastage rates used here, the annual rates of gain in the beef herd are predicted to be 0·39, 0·43, 0·43 and 0·42 × heritability × phenotypic standard deviation, with females retained for 3, 4, 5 and 6 years of breeding respectively, and having their first progeny at 2 years of age. No selection on females is practised if they are kept for only 3 years, so the selection programme is cheaper to operate, even if it is not the most efficient.

Although annual rates of progress may be computed merely from a knowledge of the generation interval, this parameter is not sufficient to specify responses in the early years of any scheme. The age distribution of parents is necessary and is shown in Table 1 for our model. These are somewhat arbitrary but will serve for illustration, and have been chosen to give typical results for the generation intervals, which are also given in the Table. In the dairy breed 30% of purebred calves are got by young dairy bulls undergoing progeny tests, but only older bulls and cows are used to breed dairy bull replacements. A higher proportion of crosses are obtained from the young cows. Beef bulls are used for only one year in the closed herd, but for four years in AI. The best beef bulls may have progeny in both the herd and the whole dairy population in the same year.

The numbers of animals in the model population are also chosen arbitrarily since our primary concern is to discuss methodology. Let us assume the dairy breed comprises 1 000 000 cows, of these 75% are mated pure and the rest are crossed. Of the purebred calves M_p = 300 000 (almost all males) are finished for beef, and of the crossbreds M_c = 200 000 are finished annually. In the dairy breed 150 bulls are required each year to enter the progeny test and 600 bulls are performance tested annually. However, we assume a wastage of 25%, so that the effective intensity is 1 in 3, and the selection differential is 1·1 standard deviations. More intense selection is unlikely to justify extra expense (Hinks, 1970b). The beef herd is maintained with 400 cows and 8 bulls. Since the generation interval is 2½ years, the rate of inbreeding is ¼% per year (Turner and Young, 1969) and should

be acceptable. Each year 16 bulls enter AI and these include the above 8. Assuming that 160 bull calves are born per year and there is a 25% wastage, the selection differentials are 1·9 and 1·6 standard deviations in the herd and in AI for crossing, respectively.

TABLE 1

Distribution (%) of parental age at birth of progeny. C = *cows (as parents or replacements)*, B = *bulls (as parents or replacements)*, SP, SX = *purebred or crossbred animals for slaughter*

		Age (years)										
		2	3	4	5	6	7	8	9	10		
Parents	Progeny	*Percentages*									Mean age	
Dairy breed												
B	B	0	0	0	0	0	0	33	33	33	9	7
B	C, SP	30	0	0	0	0	19	18	17	16	6·5	
C	B	0	0	0	0	0	40	30	20	10	8	
C	C, SP	9	28	21	15	11	8	5	3	0	4·5	
C	SX	19	29	20	14	8	5	3	2	0		4
Beef breed												
B	B, C	100	0	0	0	0	0	0	0	0	2	2·5
C	B, C	33	33	33	0	0	0	0	0	0	3	
B	SX	25	25	25	25	0	0	0	0	0		3·5

In the model, testing facilities are established in year 0 and selection is first practised in individuals born in year 1. The selected trait in both breeds is taken to be simply individual weight at 400 days of age, which is assumed to be additive, with no interaction between breeds or sexes, and with the same parameters in each breed. These are: a heritability of 0·4, a phenotypic standard deviation of 40 kg, and a partial regression of net income over feed costs on weight of £0·15 per kg. Some of these numerical values adopted are taken from Hinks (1970b). The heritability value chosen allows for some interaction between test and commercial environment and finishing age.

Summary of parameter values

For reference all the parameters (and trial values) are defined below:

d = discount rate (8%, 15% and 20%).

$r = \dfrac{1}{1+d}$ = discount factor.

t = year of obtaining return or incurring cost relative to a base year, $t = 0$.

y = year when first returns are obtained.

T = total number of years over which scheme is evaluated.

i = selection differential in standard units of index on which animals are selected (index = weight at 400 days).

σ = phenotypic standard deviation of index (for 400-day weight, σ = 40 kg).

h^2 = heritability of index or, strictly, the regression of economic genetic merit on index value (0·4 for individual 400-day weight).
g = generation interval in years.
a = net monetary return from one unit difference of performance on one animal (for 1 kg extra 400-day weight, a = £0·15).
M = number of animals slaughtered annually for beef in specified scheme (beef × dairy: M_c = 200 000; pure dairy: M_p = 300 000).
R = annual increment in monetary return from one year's selection.
C = annual testing costs.
I = initial investment in testing scheme.
I' = part of initial investment, from sale of land and stock, for example, which can be realized if the scheme is terminated.

Predicted genetic improvement

The expected changes in the purebred and crossbred populations resulting from selection in the two schemes are shown in Figure 1. The initial response is erratic as genes from selected individuals become distributed through the population. Eventually the rate of advance stabilizes at $ih^2\sigma/2g$, since selection is practised in one sex and g is the generation interval. The increment in response between successive years, say 6 and 7, can also be viewed as the increment occurring at year 7 from a single selection practised at year 1. This approaches $ih^2\sigma/2g$ as t increases. The advance in beef crosses are shown in two ways, both including and excluding selection among the bulls taken for AI. A performance test for bulls in AI could be run without maintaining a closed beef herd, so the gain made using the herd alone should not include this increment. Thus we find that no advance is made in crossbred populations until year 5, when the first grand-progeny of selected animals are born. The much more rapid advance from selection in the beef breed is also very clear.

In constructing Figure 1 the assumption has been made that the heritabilities and variances do not change with selection and inbreeding. However, the period of 15 years during which selection has an impact on returns evaluated over 20 years represents only six generations of selection and a total inbreeding of 4% in the beef breed, which is smaller and has a shorter generation interval than the dairy breed. Therefore the assumption of constant genetic parameters may not be unreasonable.

Costs and returns

We have not yet considered the costs of running the scheme. Imagine that in the dairy breed an initial investment of about £50 000 (i.e. £80 per bull for 600 bulls) is required for a performance testing house, assumed to be built on an existing bull rearing station. There would be an increase in annual expenditure of £250 for the purchase of each of 450 bulls, which would include the necessary recording and analysis of the appropriate records of potential dams. The testing costs would be about £40 for each of 600 bulls for administration, diets, etc. The increased maintenance costs would be partly offset by the slaughter value of unselected animals, the net deficiency is assumed to be £50 for each of 450 animals. The increase in annual costs therefore totals about £160 000, roughly in line with Hinks' (1970b) figures. In the herd for the beef breed we assume a farm is purchased and test facilities built on it. For a 400-cow herd a 600-acre farm

would be required, costing £150 000 for land at £250 per acre. The testing station for 160 animals per year at £80 per animal would cost £12 800. The initial purchase of stock would involve 16 bulls at £800 each, and 400 heifers or cows at £150 each, £72 800 in all. Thus the total outlay would be

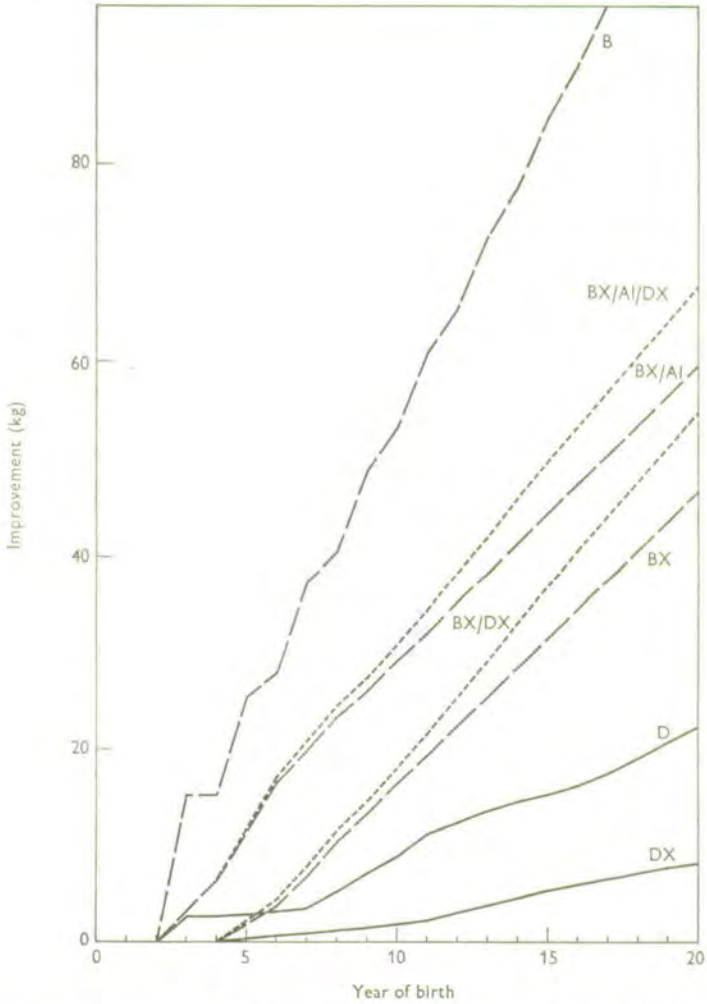

FIG. 1. Mean improvement in 400-day weight from selection in specified scheme on purebred and crossbred performance. B = pure beef breed, BX = beef cross, BX/AI = beef cross plus selection of bulls for AI, D = pure dairy breed, DX = dairy cross, BX/DX = beef plus dairy cross, BX/AI/DX = beef cross, plus selection for AI, plus dairy cross.

about £235 000. Of this, all but the cost of the test facility could be realized at any time so $I' =$ £222 000. The running costs would include £40 for each animal on test, and the rest would be covered by sale of stock, AI fees, etc., since no rent charge is being made. The annual costs are therefore put at £7000.

The returns from the schemes which will be included are derived solely from genetic improvement in the stock reared and slaughtered for beef. No returns will be counted from crossbred or purebred males or females used

for breeding in single suckler herds, nor from improvement of the selected beef herd itself, since it is very small. It is assumed that no extra maintenance cost is incurred in maintaining larger breeding or milking animals, or at least that it is balanced by the extra slaughter value of culls. In the dairy breed the rate of advance in milk production from the progeny testing scheme is not affected by the selection for beef characteristics; there is thus an implicit assumption that the correlation of growth rate and milk production is zero, and that there is no reduction in the selection intensity

TABLE 2

Discounted cash flow analysis with 8% discount rate (*All revenues are in* £'000)

		Breed selected										
		Dairy						Beef				
			Returns			Discounted			Returns		Discounted	
Year	Discount factor	Costs (−)	Pure	Cross	Net	Net	Sum	Costs (−)	Cross	Net	Net	Sum
0	1·000	50	0	0	−50	−50	−50	235	0	−235	−235	−235
1	0·926	160	0	0	−160	−148	−198	7	0	−7	−6	−241
2	0·857	160	0	0	−160	−137	−335	7	0	−7	−6	−247
3	0·794	160	0	0	−160	−127	−462	7	0	−7	−6	−253
4	0·735	160	0	0	−160	−118	−580	7	0	−7	−5	−258
5	0·681	160	119	0	−41	−28	−608	7	0	−7	−5	−263
6	0·630	160	119	0	−41	−26	−634	7	0	−7	−4	−267
7	0·584	160	124	8	−28	−16	−650	7	57	+50	+29	−238
8	0·540	160	141	19	0	0	−650	7	114	107	58	−180
9	0·500	160	154	27	+21	+11	−639	7	209	202	101	−79
10	0·463	160	239	34	113	52	−587	7	313	306	142	+63
11	0·429	160	320	40	201	86	−501	7	396	389	167	230
12	0·397	160	400	50	290	115	−386	7	491	484	192	422
13	0·368	160	505	67	412	151	−235	7	579	572	210	632
14	0·340	160	556	87	483	164	−71	7	673	666	227	859
15	0·315	160	611	112	563	177	+106	7	762	755	238	1097
16	0·292	160	650	137	627	183	289	7	855	848	248	1345
17	0·270	160	687	160	687	186	475	7	945	938	254	1599
18	0·250	160	733	180	753	188	663	7	1037	1030	258	1857
19	0·232	160	791	198	829	192	855	7	1127	1120	259	2116
20	0·214	160	861	214	915	196	1051	7	1441†	1434	308	2424

† Includes sale of farm and stock.

practised for milk production when nominating dams to breed replacement bulls.

The annual rates of genetic improvement after they have reached a steady value (see Figure 1) are 1·26 kg/year in the dairy breed and 6·08 kg/year in the beef breed. With 300 000 pure dairy and 200 000 crossbred animals slaughtered per year, these responses are worth £56 600/year in the dairy breed, and £18 900/year and £91 200/year in the crosses from selection in the dairy and beef breeds, respectively.

The detailed discounted cash flow analysis is shown in Table 2, in which a discount rate of 8% and a 20-year evaluation period are used. Both schemes are highly profitable under these conditions. However, the Table shows clearly how long one must wait before the schemes break even, 10 years for selection in the beef breed and 15 years in the dairy breed. The

effects of alternative discount rates are shown in Figure 2, in which only the aggregate discounted cash flow is shown. The beef breed scheme shows a rate of return of about 27% and the dairy breed scheme a rate of about

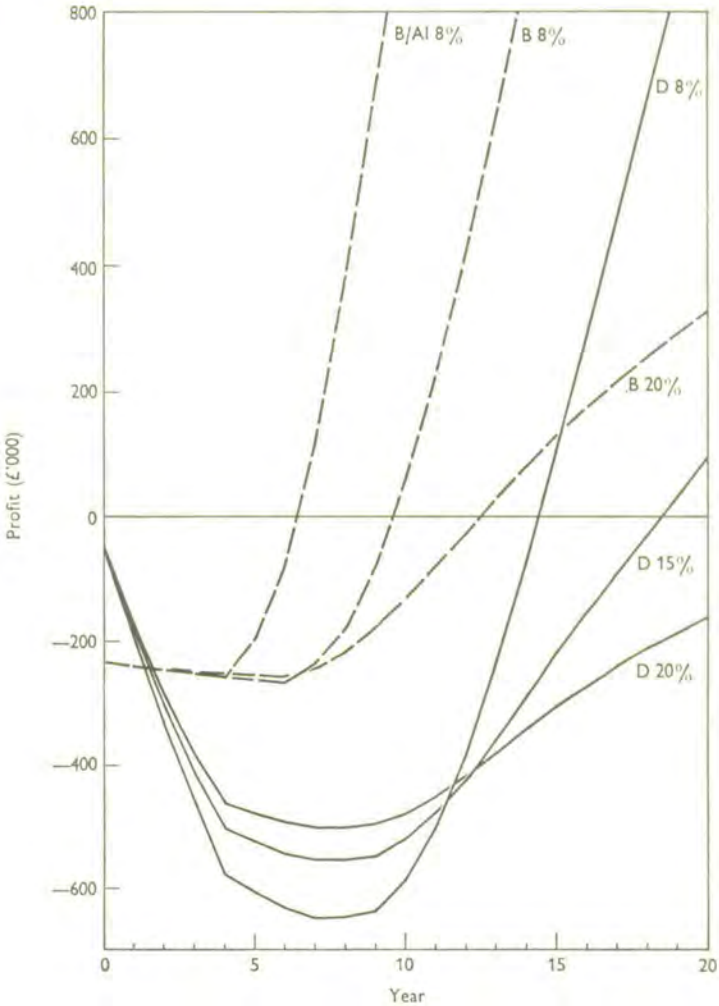

FIG 2. Aggregate net profit with discount rates shown and selection in either the beef breed (B), with AI or without selection of bulls for AI, or the dairy breed (D).

16%. With these discount rates the schemes just break even by year 20. They break even at year 15 with discount rates of about 25% and 10% respectively. The effect of the selection of bulls only used in AI is also shown in Figure 2. If this scheme is started at the same time, the rate of return is very markedly increased since gains are made in the early years when the discount factor is still close to unity, and no extra costs are involved.

ALGEBRAIC SOLUTIONS

We need to consider the effect of changes in some of the numerical values chosen in our examples on the net returns from the schemes. This

will be more feasible if we have some fairly reliable, but simple, formulae for the discounted profit in a breeding programme. Within the framework of our model such a formulation is straightforward except in taking account of the uneven rate of selection advance in the early years of the scheme. In a species such as poultry, with discrete yearly generations, all commercial stock used in a single year have the same expected performance and the problem is simpler. We formally consider this case, but show that reasonable approximations can be made to other schemes, such as those for beef cattle.

The initial investment in the programme is I (monetary units) of which a part I' (before discounting), such as a farm or stock, can be realized, and the scheme is evaluated over a period of T years. The annual testing cost is C which we have taken to commence at year 1, although the formulae can easily be modified to allow testing in year 0. The total testing expenditure, discounted to the base year, is $C(r+r^2+\ldots+r^T) = Cr(1-r^T)/(1-r)$. The annual increment is undiscounted returns from improvement of the stock is $R = Mih^2a\sigma/g$, where the parameters are defined above, and i is averaged over the two sexes. The first commercial stock is marketed at year y, i.e. there are y years of multiplication and rearing between selection in nucleus herds and obtaining monetary returns. Selection at year 1 realizes aggregate discounted returns of $R(r^y+r^{y+1}+\ldots+r^T)$, that at year 2 realizes $R(r^{y+1}+\ldots+r^T)$, and so on, giving

$$R[r^y+2r^{y+1}+\ldots+(T-y+1)r^T] = R\left[\frac{r^y-r^{T+1}}{(1-r)^2}-\frac{(T-y+1)\,r^{T+1}}{1-r}\right].$$

The total 'profit' from the scheme, P, is the sum of discounted returns less discounted costs; we have

$$P = R\left[\frac{r^y-r^{T+1}}{(1-r)^2}-\frac{(T-y+1)r^{T+1}}{1-r}\right]-\frac{Cr(1-r^T)}{1-r}-I+I'r^T. \qquad (1)$$

The 'cash flows' in the scheme are $-I$ in year 0, $-C$ in years 1 to $y-1$, $(t-y+1)R-C$ in years y to $T-1$, and $(T-y+1)R-C+I'$ at year T. These would be increased in later years, as would P, if testing were ended at year $T-y$ but returns taken until year T. Equation (1) simplifies considerably if returns are evaluated over an infinite time period. Then

$$P = Rr^y/(1-r)^2-Cr/(1-r)-I. \qquad (2)$$

Since $r = \dfrac{1}{1+d}$, equation (2) may be written in terms of the discount rate

$$P = R/[d^2(1+d)^{y-2}]-C/d-I.$$

In our beef cattle examples where the initial returns are uneven we shall approximate the schemes by assuming returns are regular from the year the first descendants of selected animals are marketed. This could lead to an over- or underestimate of total returns depending on the pattern of response. The approximations are likely to be useful in these examples, for we see in Figure 1 that the responses do not depart far from linearity. For the closed beef herd scheme the first returns are obtained when the grand-progeny of selected bulls are marketed, which occurs at year $y = 7$ when the initial testing is in year 1. (We exclude here returns from selecting the bulls to go into the AI stud.) The necessary values are: I = £235 000, I' = £220 000,

C = £7000, R = £91 000 (see above). In the dairy breed selection scheme I = £50 000, I_T = £0, C = £160 000 and y = 5 years and R_p = £56 600 for purebreds, and y = 7 years and R_c = £18 900 for crossbreds. The approximate and exact results are compared in Table 3 for an evaluation

TABLE 3

Returns (£'000) and break even discount rate for exact and approximate evaluation of model selection schemes over 20 years. First returns are obtained at $y = 5$ *for the approximation in the beef breed, and* $y_p = 5$, $y_c = 7$, *for approximation (1),* $y_p = 6$, $y_c = 8$ *for approximation (2) in the dairy breed*

		Discount rate			
Breed selected	Method	8%	15%	20%	Break-even
Beef	Exact	2424	809	328	28%
	Approx.	2715	945	415	29%
Dairy	Exact	1051	96	−163	16%
	Approx. (1)	1534	306	− 33	19%
	Approx. (2)	1069	67	−191	16%

period of T = 20 years, using discount rates of 8%, 15% and 20%. The break even discount rate (or rate of return) is also computed. The agreement is very good for the beef breed and any discrepancy is probably of small magnitude relative to the changes in returns obtained by altering some of the other assumptions. For the dairy breed the approximations are poorer if the actual times when first returns are obtained are used ($y_p = 5$, $y_c = 7$). Much better approximations are obtained with $y_p = 6$, $y_c = 8$ and these values are used in the later development. Thus we do not need to be too concerned about the uneven rate of initial advance to selection and consequent returns, although the approximations would be less satisfactory if a smaller value of T were adopted.

Modification of assumptions

The effect of changes in the values assumed for the parameters can be evaluated either by the difference in profit for a fixed discount rate, or by the difference in rate of return (discount rate for zero profit). The latter may be more meaningful but requires greater numerical work for evaluation. We shall consider both alternatives.

For a fixed period of T = 20 years and the appropriate y values, equation (1) reduces to the following equations for the beef and dairy breed selection schemes:

Breed selected	Discount rate		
Beef	8%	$P = 32{\cdot}58R - 9{\cdot}82C - I + 0{\cdot}21I'$	(3)
	15%	$P = 13{\cdot}27R - 6{\cdot}26C - I + 0{\cdot}06I'$	
	20%	$P = 7{\cdot}44R - 4{\cdot}87C - I + 0{\cdot}03I'$	
Dairy	8%	$P = 38{\cdot}42R_p + 27{\cdot}40R_c - 9{\cdot}82C - I$	
	15%	$P = 16{\cdot}18R_p + 10{\cdot}80R_c - 6{\cdot}26C - I$	
	20%	$P = 9{\cdot}32R_p + 5{\cdot}89R_c - 4{\cdot}87C - I$	

The differentials of P with respect to R, C and I are obtained immediately from equations (3); we notice that the value of R is most important when the discount rate is small. In our model $R = Mih^2a\sigma/g$, so a specified proportional change in M, i, h^2, a, or σ gives the same change in profit. For example, from (3)

$$\frac{\partial P}{\partial a/a} = \frac{\partial P}{\partial M/M} = 32{\cdot}58R$$

for the beef breed using an 8% discount rate. Thus a 1% increase in the estimate of a or M increases the prediction of P by £0·01 × 32·58 × 91 200 = £29 000, by no means a negligible sum of money. The differentials of profit on annual costs (C) are the same for the beef and dairy schemes, but the estimates of C are £7000 and £160 000, respectively, so that $\partial P/(\partial C/C)$ is about 23 times as large for the dairy as for the beef breed. A proportionate increase in costs is much more serious in the dairy breed.

An alternative approach is to examine the value of return (R) necessary for the scheme just to break even. Again let us assume that $R_c = \frac{1}{3}R_p$ for the dairy breed. Using equations (3) and making the previous assumptions for C, I and I', we find the break even values are R = £7900 and £35 300 for the beef breed, and $R_p + R_c$ = £45 400 and £87 000 for the dairy breed with $d = 8\%$ and 20% respectively. These values for returns are much closer to those assumed earlier in the dairy breed than in the beef breed. It is clear that there could be considerable error of estimation of the parameters in the beef breed, yet the scheme would still be profitable.

Finally we can investigate the effect on the discounted yield (d) of changes in assumption. Numerical solution is now necessary, and since the regression of d on the parameter values is non-linear the effect of both small changes (10% which approximates the derivative) and large changes (halving and doubling) of the values have been considered. The results are summarized in Table 4, and the effects of changes in y and T, the number of years to first

TABLE 4

Effect of changes in assumptions, one at a time, on the rate of return (d%), *for beef breed and dairy breed schemes. Base values in £'000 or years are* R = 91·2, C = 7, I = 235, I′ = 222, y = 7, T = 20 *for the beef breed, giving* d = 29·42%; *and* R_p = 56·6, R_c = 18·9, C = 160, I = 50, I′ = 0, y_p = 6, y_c = 8 *for the dairy breed, giving* d = 15·96%.

Variable changed	Beef breed			Dairy breed		
	Amount of change			Amount of change		
	×0·5	×1·1	×2	×0·5	×1·1	×2
$M, i, h^2, a, \sigma, 1/g$ } $R = R_p + R_c$	22·38	30·49	37·30	4·89	17·43	26·81
C	29·92	29·22	28·46	25·65	14·56	5·11
I	36·43	28·44	22·95	16·36	15·89	15·24
Amount (years)	−1	+1		−1	+1	
y, y_p, y_c	32·75	26·58		19·32	13·00	
Amount (years)	−5	+5		−5	+5	
	26·83	30·19		9·36	18·25	

returns and the total evaluation period, respectively, are shown. From these results it is again apparent that considerable errors can be made in the estimates of the parameters for the beef breed scheme, yet a high discount rate can still be achieved. The discounted return from the dairy breed scheme is rather more sensitive to some of the assumptions. The parameter which is most difficult to predict is M, the total number marketed in the scheme, for the beef breed programme. There may be no prior guarantees about numbers of crosses of the beef herd with the dairy breed, whereas the 'market' for the breeding programme in the dairy breed is more assured.

DISCUSSION

Although there are many implicit assumptions in the methods of investment appraisal which have been described, they may help in reaching rational decisions about proposed breeding schemes. In the examples considered in this paper we see that the beef breeding scheme could be highly cost-effective, and the dairy breed scheme less so, but nevertheless be profitable at interest rates of up to 15%. There remain many limitations in the approach. For example all the parameters have been assumed to remain constant, yet heritability may change with inbreeding and selection, and in particular the returns from the schemes will rise or fall in line with total beef consumption and will be influenced by the success of competitive schemes. Some simple predicted changes in parameters can easily be included, even in the algebraic formulation. For example if M, the number marketed, increases by a proportion k each year, i.e. M, Mk, Mk^2, . . ., then the effective discount factor for returns is simply rk and equation (1) can readily be modified. But generally these changes in parameters are difficult to predict with any accuracy at all. Since there exists these large elements of uncertainty new schemes should be evaluated over short time periods (T) with high discount rates. Perhaps the time period used here (20 years) is too long, and a period of only 15 years would be justified. However, 6–8 years may be required before any returns are obtained from the scheme, so for T much below 15 years none of the beef programmes will appear profitable.

The analysis here has been made for new breeding programmes which require capital expenditure before they can be started. When they are in operation the appraisal procedure can be modified. The important decision then may be whether another year of selection should be practised. This will be justified if the predicted returns from the *one* selection are adequate to cover costs at the required yield. Hinks (1970a, b) discussed the efficiency of programmes for improving milk yield and beef production and in a dual purpose breed using such an approach. When the initial capital costs are not large, or the testing facilities already exist the 'one selection' approach is perhaps more relevant. If the commercial benefits of a single year's selection are first realized y years later, and are assumed to remain constant, the total discounted returns obtained over a T year period from a single selection are $R(r^y - r^{T+1})/(1-r)$.

No attempt has been made here to use the discounted cash flow method of estimating returns for finding the optimal design of the programme. Hinks (1970a, b) considered the selection intensity which should be practised, and the algebraic results obtained here can be used for this purpose. The

returns are proportional to i, the standardized selection differential. If the total number of animals selected is fixed, $1/p$ animals are recorded for every one selected, if p is the fraction selected. Then the initial capital and annual testing costs can be assumed to be proportional to $1/p$. At the discount rate which the scheme is hoped to realize, let α be the aggregated discounted returns per standard deviation of selection differential, and let β be the sum of the aggregated discounted animal costs and initial capital expenditure, per animal selected, such that the total profit is $P = \alpha i - \beta/p$. The optimal selection intensity is given by $\partial P/\partial p = 0$. A good approximation to i for normally distributed populations is given by Smith (1969):

$$i = 0{\cdot}8 + 0{\cdot}41\ ln(1/p - 1).$$

If only a proportion, s, of animals are available for selection, since some die or are unfit for breeding, Smith's formula becomes

$$i = 0{\cdot}8 + 0{\cdot}41\ ln\ (s/p - 1)$$

and, at the optimum, we obtain

$$p = \frac{\beta}{\beta + 0{\cdot}41 s\alpha}. \tag{4}$$

Consider the beef breed example, assuming a 20% discount rate. Our original assumptions were $p = 0{\cdot}05$, $s = 0{\cdot}75$ and $i = 1{\cdot}9$. Working back from equation (3), we obtain $\alpha = (91\,200 \times 7{\cdot}44)/1{\cdot}9 = £357\,000$ and similarly $\beta = £13\,000$. Substituting into (4), we have at the optimum, $p = 0{\cdot}11$, indicating that the proposed selection is too intense. But with a discount rate of 8%, the optimal intensity is $p = 0{\cdot}026$, since the monetary returns are so much larger for each unit of improvement.

There are clearly many other problems raised in this analysis, particularly since the conclusions may be so dependent on the discount rate demanded of the scheme. Further, no account has been taken of any benefits which may accrue from the scheme in years after the returns have been discounted to zero. Many breeding programmes have an impact for periods much longer than 20 years.

ACKNOWLEDGEMENTS

I have had the benefit of many valuable discussions with Dr C. J. M. Hinks and Mr R. C. Rickard, who, together with Dr L. K. O'Connor, Dr D. M. Allen and Mr D. E. Steane, provided useful numerical information on which the models are based. Miss J. Carne gave considerable technical assistance.

REFERENCES

HINKS, C. J. M. 1970a. The selection of dairy bulls for AI. *Anim. Prod.* **12**: 569–576.

HINKS, C. J. M. 1970b. Performance test procedures for meat production amongst dairy bulls used in AI. *Anim. Prod.* **12**: 577–583.

HOUSE OF COMMONS (1967–8). *Select Committee on Nationalised Industries.* 1st report, Pp. 1–22 (H.C. 371).

LINDHÉ, B. 1968. Model simulation of A.I. breeding within a dual purpose breed of cattle. *Acta Agric. scand.* **18**: 33–41.

POUTOUS, M. and VISSAC, B. 1962. Recherche théorique des conditions de rentabilité maximum de l'épreuve de descendance des taureaux d'insémination artificielle. *Ann. Zootech.* **11**: 233–256.

SEARLE, S. R. 1961. Estimating herd improvement from selection programs. *J. Dairy Sci.* **44**: 1103–1112.

SMITH, C. 1969. Optimum selection procedures in animal breeding. *Anim. Prod.* **11**: 433–442.

SOLLER, M., BAR-ANAN, R. and PASTERNAK, H. 1966. Selection of dairy cattle for growth rate and milk production. *Anim. Prod.* **8**: 109–119.

TURNER, H. N. and YOUNG, S. S. Y. 1969. *Quantitative Genetics and Sheep Breeding.* Macmillan, Melbourne.

(*Received* 15 *July* 1970)

17

Prediction and evaluation of response to selection with overlapping generations

by

William G. Hill

Anim. Prod. 1974, **18**: 117–139

PREDICTION AND EVALUATION OF RESPONSE TO SELECTION WITH OVERLAPPING GENERATIONS

WILLIAM G. HILL

Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN

SUMMARY

In a population in which generations overlap the improvement in performance in successive years resulting from a single year of selection is not constant, for the genes from a group of selected individuals may take many years to pass through the population. A formal method is developed for predicting responses and discounted returns from improvement in populations with overlapping generations including, if necessary, generations of multiplication of breeding stock.

The method is based on a matrix which specifies the passage of genes between the different age groups and sexes. Simple matrix operations can be used to compute the proportion of genes in animals of both sexes and each age in the population at any time which derive from a group of selected animals at an earlier time. The response produced by these selected animals equals the product of their genetic selection differential and the proportion of genes deriving from them.

Comparisons are made between responses predicted using this theory and the classical theory of uniform rates of response, and a method is given for computing the time lag of genes passing through the population.

The results are extended to enable computation of discounted returns from improvement.

INTRODUCTION

THE classical theory of response to artificial selection in populations in which generations overlap was developed by Dickerson and Hazel (1944) and Rendel and Robertson (1950). It enables prediction of the rate of response when the same selection scheme is practised for many generations; but, when generations overlap, the genetic improvement in the selected group of animals in one year is not immediately passed through the population, as it is if generations are discrete. For example, in a dairy cattle population selected bulls may only be used for a year or two, yet some of their progeny live for 10 years or more. Thus the effect of a single cycle of selection on the performance of subsequent generations is erratic for many years after the selection is practised, as Hinks (1971) and Hill (1971) have pointed out. The rates of response predicted by the classical theory are therefore reached only asymptotically. Alternative methods of computing the progress each year as a result of selection in a population with overlapping generations have been given by Searle (1961), Poutous and Vissac (1962), Van Vleck (1964), Hinks (1970, 1971, 1972) and Hill (1971).

The aim of the present paper is to present a formal procedure for predicting response with overlapping generations, using matrix methods, which has several applications: it gives an insight into the genetic structure of a population with overlapping generations; enables prediction of short-term response and shows the relation of this to the asymptotic rate; provides a simple method for computing the time lag (Bichard, 1971) of improvement from nucleus to commercial stock; and allows computation of discounted monetary returns from a breeding programme, which depend most on responses in early years. Some of the methods described have been used less formally previously (Hill, 1971; A. Robertson, personal communication) and more formally as part of an analysis of effective size of populations with overlapping generations (Hill, 1972). Whilst the methods do not enable us to compute results which cannot be obtained in other ways, such as those of Hinks (1971, 1972), they considerably simplify the analysis, provide a general solution and enable standard computer routines for matrix operations to be used. The basic structure and some of the matrix results have recently been obtained independently by J. M. Elsen (personal communication).

The plan of the paper is as follows: firstly the parameters necessary to describe the structure of a population with overlapping generations are defined and used to show what proportion of genes deriving from a particular group of animals of specified age and sex are present in animals born in subsequent generations. These results are used to predict the response expected from a single group of selected animals, and to show that the response obtained many generations later equals the rate of response predicted by classical theory. The response to continued selection, and its departure from that predicted by a uniform rate of response are then considered. These main results are extended in several ways: to include programmes of multiplication of breeding stock, to take account of different selection intensities for parents chosen to breed male versus female replacements (as in dairy cattle) and to enable prediction of monetary returns. The longer mathematical proofs are included in an Appendix.

PREDICTION OF RESPONSE

Source of genes

As a basis for computations of rates of progress it is useful to find what proportion of genes of individuals born in each successive year (or specified time period) derive from the group of males or females born in some particular reference year. The age distribution of parents of newborn individuals is assumed to remain the same each year. The methods will be illustrated with a very simple example which was also used by Bichard, Pease, Swales and Ozkutuk (1973). The matrix equations hold generally.

For the example, consider a population of pigs in which there is a discrete farrowing period every 6 months. Boars are used in only one mating season, and have their progeny when 12 months old. Sows farrow twice, at 12 and 18 months, and have an equal number of progeny each time. It is convenient to take 6-month time periods, so that boars are 2 time units and sows 2 and 3 time units old when their progeny are born. Thus animals born at time t obtain one-half of their genes from males aged 2 units, one-quarter from females aged 2 and one-quarter from females aged 3 units at that time. Males

of age 1 unit at time t obtain all their genes from males aged 0 units at $t-1$ since they are the same individuals.

The passage of genes in the population can be expressed as follows. Let $M(i, t)$, $F(i, t)$ be the proportion of genes in males, females of age i units at time t which derive from some specified group of animals at time 0. We have

$$M(0, t) = \tfrac{1}{2}M(2, t)+\tfrac{1}{4}F(2, t)+\tfrac{1}{4}F(3, t), \tag{1a}$$

$$M(1, t) = M(0, t-1),\ M(2, t) = M(1, t-1), \tag{1b}$$

and

$$F(0, t) = \tfrac{1}{2}M(2, t)+\tfrac{1}{4}F(2, t)+\tfrac{1}{4}F(3, t), \tag{1c}$$

$$F(1, t) = F(0, t-1),\ F(2, t) = F(1, t-1),\ F(3, t) = F(2, t-1). \tag{1d}$$

Using these equations we obtain

$$M(1, t) = \tfrac{1}{2}M(2, t-1)+\tfrac{1}{4}F(2, t-1)+\tfrac{1}{4}F(3, t-1), \tag{1e}$$

$$F(1, t) = \tfrac{1}{2}M(2, t-1)+\tfrac{1}{4}F(2, t-1)+\tfrac{1}{4}F(3, t-1). \tag{1f}$$

Since equations (1b), (1d), (1e) and (1f) now completely specify one time period in terms of the previous one, an iterative process can be used to compute the proportion of genes in animals at time t deriving from the group at time 0.

These equations can alternatively be put into matrix terms. Let $\boldsymbol{m}_{(t)}$ be a vector with elements $m_{i(t)}$ where

$$m_{1(t)} = M(1, t),\ m_{2(t)} = M(2, t),\ m_{3(t)} = F(1, t),\ m_{4(t)} = F(2, t),\ m_{5(t)} = F(3, t).$$

For example, (1e) becomes

$$m_{1(t)} = \tfrac{1}{2}m_{2(t-1)}+\tfrac{1}{4}m_{4(t-1)}+\tfrac{1}{4}m_{5(t-1)}.$$

These equations in matrix notation are

$$\boldsymbol{m}_{(t)} = \boldsymbol{P}\boldsymbol{m}_{(t-1)} \tag{2}$$

where

$$\boldsymbol{P} = \left(\begin{array}{cc|ccc} 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ 1 & 0 & 0 & 0 & 0 \\ \hline 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right). \tag{3}$$

The blocks of $\boldsymbol{P}$ correspond to the alternative pathways of genes

$$\left(\begin{array}{c|c} \text{males to males} & \text{females to males} \\ \hline \text{males to females} & \text{females to females} \end{array}\right).$$

The elements p_{ij} of $\boldsymbol{P}$ are defined as the proportion of genes in animals of sex-age class i at time t coming from animals of sex-age class j at time $t-1$. The general form of the matrix is

$$\boldsymbol{P} = \left[\begin{array}{cccc|ccc} p_{11} & p_{12} & \cdots & p_{1,h-1} & p_{1,h} & p_{1,h+1} & \cdots & p_{1,h+k} \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline p_{h+1,1} & p_{h+1,2} & \cdots & p_{h+1,h-1} & p_{h+1,h} & p_{h+1,h+1} & \cdots & p_{h+1,h+k} \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{array}\right]$$

The dimensions of the blocks, h and k, have to be large enough to include all breeding animals, and $\boldsymbol{P}$ is square of dimension $h+k$. The off-diagonal elements of unity exhibit the passage of genes due to ageing, those in the first row of each block the passage due to reproduction. Since one-half of the genes come from parents of each sex

$$\sum_{j=1}^{h} p_{1j} = \sum_{j=h+1}^{h+k} p_{1j} = \sum_{j=1}^{h} p_{h+1,j} = \sum_{j=h+1}^{h+k} p_{h+1,j} = 0{\cdot}5;$$

and, for $i = 2, \ldots, h, h+2, \ldots, h+k$, $p_{i,i-1} = 1, p_{ij} = 0, j \neq i-1$.

Note that all row totals of $\boldsymbol{P}$ equal unity.

The matrix $\boldsymbol{P}$ is an extension of the type defined by Leslie (1945) for studies of population growth which has subsequently been used by many other authors. In that situation, the matrix refers only to the total population number or number of females present, there is no separate matrix partition for males and females and neither row nor column totals necessarily equal unity. In the analysis in this paper the minimum age included in the matrix is taken as 1 time unit. An alternative is to take it as 0 time units by including newborn individuals, but then parental ages in the matrix appear as 1 time unit before their progeny are born. The analysis can be done either way.

As one specific example, let $\boldsymbol{m}_{(t)}$ be the proportion of genes in individuals at time t deriving from males of age 1 unit at time 0. Then the transpose of $\boldsymbol{m}_{(0)}$, denoted $\boldsymbol{m}'_{(0)}$ (used to reduce space), is

$$\boldsymbol{m}'_{(0)} = (1 \quad 0 \mid 0 \quad 0 \quad 0).$$

As a second example, let $\boldsymbol{f}_{(t)}$ be the proportion of genes coming from females of age 1 at time 0,

$$\boldsymbol{f}'_{(0)} = (0 \quad 0 \mid 1 \quad 0 \quad 0)$$

and $\boldsymbol{f}_{(t)} = \boldsymbol{P}\boldsymbol{f}_{(t-1)}$.

In Table 1 the elements of $\boldsymbol{m}_{(t)}$ and $\boldsymbol{f}_{(t)}$ are tabulated for several successive time periods. The values can be obtained using (2) repeatedly, or from

$$\boldsymbol{m}_{(t)} = \boldsymbol{P}^t\boldsymbol{m}_{(0)}, \quad \boldsymbol{f}_{(t)} = \boldsymbol{P}^t\boldsymbol{f}_{(0)}$$

directly. The table shows that the proportions of genes deriving from a source group which are present in individuals born at different times fluctuate initially but eventually stabilize, in this case at a value of 0·222 = 2/9 when $\boldsymbol{m}_{(0)}$ and $\boldsymbol{f}_{(0)}$ contain a single value of unity as shown. An explanation of why the asymptotic value is 2/9 in this example is given subsequently.

A study of all the eigenvalues and vectors of $\boldsymbol{P}$ which should, in principle, enable a description of the rate of approach to equilibrium, has not proved rewarding.

TABLE 1

Proportions of genes deriving from males or females of age 1 *time unit at the outset, for the pig population defined by equation* (3)

	Source of genes									
	Males Element of $\boldsymbol{m}_{(t)}$					Females Element of $\boldsymbol{f}_{(t)}$				
Time	1	2	3	4	5	1	2	3	4	5
0	1·000	0·000	0·000	0·000	0·000	0·000	0·000	1·000	0·000	0·000
1	0·000	1·000	0·000	0·000	0·000	0·000	0·000	0·000	1·000	0·000
2	0·500	0·000	0·500	0·000	0·000	0·250	0·000	0·250	0·000	1·000
3	0·000	0·500	0·000	0·500	0·000	0·250	0·250	0·250	0·250	0·000
4	0·375	0·000	0·375	0·000	0·500	0·188	0·250	0·188	0·250	0·250
5	0·125	0·375	0·125	0·375	0·000	0·250	0·188	0·250	0·188	0·250
6	0·281	0·125	0·281	0·125	0·375	0·203	0·250	0·203	0·250	0·188
8	0·242	0·188	0·242	0·188	0·281	0·215	0·234	0·215	0·234	0·203
10	0·228	0·211	0·228	0·211	0·242	0·220	0·227	0·220	0·227	0·215
15	0·222	0·223	0·222	0·223	0·221	0·222	0·222	0·222	0·222	0·223
20	0·222	0·222	0·222	0·222	0·222	0·222	0·222	0·222	0·222	0·222

Contribution of genes by reproduction alone. The values given in Table 1 include genes coming both from ageing of the original group of animals, elements $m_{2(1)}$, $f_{4(1)}$ and $f_{5(2)}$, and those from reproduction of the original group. When discussing selection response in the subsequent sections, only genes from reproduction of the original (selected) group will have to be included. A simple device can be incorporated into the analysis to remove the contribution of genes by ageing. This is based on a matrix $\boldsymbol{Q}$, with elements q_{ij}, which has unit off-diagonal elements relevant to ageing, and so equals $\boldsymbol{P}$ but with the two rows, 1 and $h+1$, for reproduction set to zero, i.e.

$$q_{i,i-1} = 1,\ i = 2, \ldots, h, h+2, \ldots, h+k,$$

$$q_{ij} = 0, \text{ otherwise.}$$

In our example (3),

$$\boldsymbol{Q} = \left(\begin{array}{cc|ccc} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right)$$

and we see that with

$$f_{(0)} = \begin{bmatrix} 0 \\ 0 \\ \hline 1 \\ 0 \\ 0 \end{bmatrix}, \text{ then } Qf_{(0)} = \begin{bmatrix} 0 \\ 0 \\ \hline 0 \\ 1 \\ 0 \end{bmatrix}, \; Q^2 f_{(0)} = \begin{bmatrix} 0 \\ 0 \\ \hline 0 \\ 0 \\ 1 \end{bmatrix}$$

and $Q^t f_{(0)} = \mathbf{0}$, $t>2$. (In general $Q^t = \mathbf{0}$, $t \geqq \max(h, k)$.) The elements of $Q^t m_{(0)}$ and $Q^t f_{(0)}$ specify the passage of genes from the original group from ageing alone. Thus the passages from reproduction alone are specified by

$$m_{(t)} = (P^t - Q^t)m_{(0)}, \quad f_{(t)} = (P^t - Q^t)f_{(0)}. \tag{4}$$

The elements of these new vectors $m_{(t)}$ and $f_{(t)}$ are given as before in Table 1, except that the unit elements in generations 0, 1 and 2 are now zero.

Response to one cycle of selection

Now assume that selection is practised among young animals, such that only superior ones are kept to age 1 time unit as potential breeders. If the superiority in breeding value of these males above the mean for the whole age group is G_m, the increment in performance of animals in successive time periods due to this selection will equal the product of G_m and the proportion of genes from the selected group. Thus the increment is given by the elements of $m_{(t)}G_m$. Similarly, if females have a genetic selection differential of G_f, the increment is $f_{(t)}G_f$. The response among the individuals of different sexes and ages at time t, $r_{(t)}$ say, is then

$$\begin{aligned} r_{(t)} &= m_{(t)}G_m + f_{(t)}G_f \\ &= (P^t - Q^t)(m_{(0)}G_m + f_{(0)}G_f) \end{aligned} \tag{5}$$

using (4), since only genes passing to progeny of these animals are relevant.

To simplify (5) a vector s, given by

$$s = m_{(0)}G_m + f_{(0)}G_f, \tag{6}$$

can be defined, which specifies both the genetic selection differentials for each sex and the ages of the animals in which selection is practised. The vector s can be defined directly, without reference to $m_{(0)}$ and $f_{(0)}$, and is fundamental to the analysis. Its elements are the genetic selection differentials of animals of the specified age and sex relative to the whole contemporary age-sex group. For example, if further selection were practised among females after they had their first litter, then the additional selection differential would become the relevant (in the pig example the last) element of s. At this stage of the analysis assume that the same animals are used for breeding male and female replacements. Removal of this assumption is deferred to a subsequent section. The responses at time t from the selection practised at time 0 are, using (5) and (6),

$$r_{(t)} = (P^t - Q^t)s. \tag{7}$$

Returning to the example with selection on young animals only, assume that it is practised for live-weight gain, with a genetic selection differential in males of $G_m = 50$ g/day and in females of $G_f = 35$ g/day (corresponding

roughly to a phenotypic standard deviation of 70 g/day, a selection intensity of 1/40 in males and 1/8 in females and a heritability of 0·3). Therefore

$$s' = (50 \quad 0 \mid 35 \quad 0 \quad 0)$$

and for individuals aged 1 time unit selected at time 0 (say spring, 1970) the resulting increment of individuals aged 1 unit at time 6 (i.e. slaughtered in spring, 1973) is, using (5) and Table 1,

$$r_{1(6)} = r_{3(6)} = 0{\cdot}281 \times 50 + 0{\cdot}203 \times 35$$
$$= 14{\cdot}1 + 7{\cdot}1 = 21{\cdot}2 \text{ g/day.}$$

This and other values are given in the 'nucleus' columns in Table 2, which can be obtained from (5) or (7). For illustration, the contributions from selection in the two sexes separately are also shown.

TABLE 2

Response (live-weight gain, g/day, in animals of age 1 unit) to a single cycle of selection and to continued selection in the examples with genetic selection differentials of 50 g/day in males and 35 g/day in females. These are compared with predictions on the basis of a uniform response

	Population											
	Nucleus Single selection at time 0 Sex selected			Nucleus Continued selection Sex selected			Commercial Continued selection Sex selected			Nucleus/commercial Uniform predictions for continued selection Sex selected		
Time	Males	Females	Both	Males	Females	Both	Males	Females	Both	Males	Females	Both
1	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	11·1	7·8	18·9
2	25·0	8·8	33·8	25·0	8·8	33·8	0·0	0·0	0·0	22·2	15·6	37·8
3	0·0	8·8	8·8	25·0	17·5	42·5	25·0	0·0	25·0	33·3	23·3	56·7
4	18·8	6·6	25·3	43·8	24·1	67·8	25·0	0·0	25·0	44·4	31·1	75·6
5	6·2	8·8	15·0	50·0	32·8	82·8	43·8	4·4	48·1	55·6	38·9	94·4
6	14·1	7·1	21·2	64·1	39·9	104·0	47·9	8·8	56·7	66·7	46·7	113·3
8	12·1	7·5	19·6	85·6	55·6	141·2	71·3	19·3	70·7	88·9	62·2	151·1
10	11·4	7·7	19·1	107·5	71·3	178·8	94·2	31·8	126·0	111·1	77·8	188·9
15	11·1	7·8	18·9	163·0	110·2	273·1	150·1	67·4	217·5	166·7	116·7	283·3
20	11·1	7·8	18·9	218·5	149·1	367·6	205·6	105·4	310·9	222·2	155·6	377·8

Asymptotic response. The expected increment in response from a single selection which is achieved by individuals born many generations later is

$$\lim_{t\to\infty} r_{(t)} = \lim_{t\to\infty} (P^t - Q^t)s \tag{8}$$

from (7). Since $Q^t = 0$ when t exceeds h or k (the dimensions of the blocks of Q), only A is required, where

$$A = \lim_{t\to\infty} P^t. \tag{9}$$

First, define a vector v, of dimension $h+k$, which has elements

$$\left.\begin{aligned} v_i &= \sum_{j=i}^{h} (p_{1j} + p_{h+1,j}), \quad i = 1, \ldots, h \\ v_i &= \sum_{j=i}^{h+k} (p_{1j} + p_{h+1,j}), \quad i = h+1, \ldots, h+k \end{aligned}\right\}. \tag{10}$$

In the pig example

$$\boldsymbol{v}' = (1 \quad 1 \mid 1 \quad 1 \quad 0{\cdot}5).$$

The elements v_i are proportional to the reproductive values, or expected gene contribution, of animals of age-sex class i. Thus young animals of either sex have value 1, and in the pig example, females of age 3 units (or, conceptually, almost 3 units) a value of 0·5, since they have already had half their progeny. If age groups in $\boldsymbol{P}$ of animals after they had finished breeding had been included, their reproductive values would all have been zero. It is shown in Appendix 1 that all rows of $\boldsymbol{A}$ are the same and equal to $\boldsymbol{v}'/2L$, so

$$\boldsymbol{A} = \mathbf{1}\boldsymbol{v}'/2L \tag{11}$$

where $\mathbf{1}' = (1\ 1 \ldots 1)$. The quantity L is the generation interval or mean age of parents of new born progeny:

$$L = \tfrac{1}{2}\sum_{i=1}^{h} i(p_{1i} + p_{h+1,i}) + \tfrac{1}{2}\sum_{i=1}^{k} i(p_{1,h+i} + p_{h+1,h+i})$$

$$= (L_{mm} + L_{mf} + L_{fm} + L_{ff})/4.$$

The values L_{mm} and L_{mf} are the average age of males when their male and female progeny, respectively, are born and L_{fm}, L_{ff} are the equivalent quantities for female parents. In the example, $L_{mm} = L_{mf} = 2$, $L_{fm} = L_{ff} = 2{\cdot}5$ and $2L = 4{\cdot}5$ time units. Hence a row of $\boldsymbol{A}$ is given by

$$\boldsymbol{v}'/4{\cdot}5 = (0{\cdot}222 \quad 0{\cdot}222 \mid 0{\cdot}222 \quad 0{\cdot}222 \quad 0{\cdot}111).$$

Note that 0·222 is the value reached by $\lim_{t\to\infty} \boldsymbol{P}^t\boldsymbol{m}_{(0)}$ in Table 1.

Combining (8), (9) and (10) we find that

$$\lim_{t\to\infty} \boldsymbol{r}_{(t)} = (\boldsymbol{v}'\boldsymbol{s}/2L)\mathbf{1}$$

$$= \left(\sum_{i=1}^{h+k} v_i s_i/2L\right)\mathbf{1} = \boldsymbol{r}_{(\infty)} \tag{12}$$

i.e. a vector with each element equal to the sum of the products of genetic selection differentials and reproductive values of animals at the age of selection, divided by twice the generation interval. In the pig example, (12) shows that each element of $r_{(\infty)}$ is $0{\cdot}222 \times 50 + 0{\cdot}222 \times 35 = 18{\cdot}9$ g/day, agreeing with the result given in Table 2.

Rendel and Robertson (1950) showed that the rate of response to selection equalled the ratio of mean genetic selection differential to mean generation interval. In the terminology of this paper the mean selection differential is $\Sigma v_i s_i/2$, which is equivalent to Rendel and Robertson's for the case where the same selection differentials are applied to breeders of males and females. (The generalization to remove this latter restriction is given in a subsequent section.) Therefore it has been demonstrated that the long-term response from selection in a single time unit is equal to the asymptotic rate of response in a continuing programme, as Hinks (1971) and Hill (1971) have argued. The final values, $\boldsymbol{r}_{(\infty)}$, given in Table 2, 18·9 g/day, are therefore the rates of response predicted by classical theory.

Some clarification and an alternative formulation of (12) may be useful. The genetic selection differentials have been defined to be those applied at a particular age, with the assumption that the same individuals are retained to later ages; therefore after the initial selection (e.g. on performance test) all other entries in $\boldsymbol{s}$ are zero unless further selection is applied. An alternative approach is to define a vector, say $\boldsymbol{s}^*$, defining the cumulative selection differential to that age,

$$\boldsymbol{s}^{*\prime} = (50 \quad 50 \mid 35 \quad 35 \quad 35)$$

and a vector, say $\boldsymbol{v}^*$, giving the contribution of genes by reproduction of individuals of each age. Therefore $\boldsymbol{v}^*$ is the sum of rows 1 and $h+1$ of $\boldsymbol{P}$, and from (3)

$$\boldsymbol{v}^{*\prime} = (0 \quad 1 \mid 0 \quad \tfrac{1}{2} \quad \tfrac{1}{2}).$$

The asymptotic response is obtained from the equivalent formulae

$$\boldsymbol{r}_{(\infty)} = (\boldsymbol{v}'\boldsymbol{s}/2L)\mathbf{1} = (\boldsymbol{v}^{*\prime}\boldsymbol{s}^*/2L)\mathbf{1}.$$

Response to repeated selection

Now consider a programme in which the same selection procedure is practised on each successive group of animals born, and in which the genetic parameters are assumed to remain constant. The response at time t from selection at time 1 is equal to that at time $t-1$ from selection at time 0, and so on. Thus the total response up to time t, expressed by the vector $\boldsymbol{R}_{(t)}$ for animals of the alternative sexes and ages, is

$$\begin{aligned}\boldsymbol{R}_{(t)} &= \boldsymbol{r}_{(t)} + \boldsymbol{r}_{(t-1)} + \ldots + \boldsymbol{r}_{(0)} \\ &= [(\boldsymbol{I} + \boldsymbol{P} + \boldsymbol{P}^2 + \ldots + \boldsymbol{P}^t) - (\boldsymbol{I} + \boldsymbol{Q} + \boldsymbol{Q}^2 + \ldots + \boldsymbol{Q}^t)]\boldsymbol{s} \qquad (13)\end{aligned}$$

using (7), where $\boldsymbol{I}$ is the identity matrix. The cumulative responses for the pig example are shown in Table 2 and denoted 'nucleus'. For comparison, the responses predicted using a uniform rate of response from (12) are also given in the Table. The difference between the predictions of the exact method and the approximation from the uniform rate of response is initially variable and sometimes large but after a few generations there remains only a constant difference of 10·2 g/day. As Hill (1971) and Hinks (1971) have shown, the departures from predictions are larger when animals are retained for many more breeding seasons, such as in cattle.

The matrix analysis can be developed further to modify (13) into a form which exhibits the departure of the response from the uniform prediction. Let

$$\boldsymbol{B} = \boldsymbol{P} - \boldsymbol{A},$$

where $\boldsymbol{A}$ is given by (9) and (11). From Appendix 2,

$$\boldsymbol{P}^t = \boldsymbol{A} + \boldsymbol{B}^t, \quad t > 0.$$

Thus (13) can be rewritten

$$\boldsymbol{R}_{(t)} = [t\boldsymbol{A} + (\boldsymbol{I} - \boldsymbol{B}^{t+1})(\boldsymbol{I} - \boldsymbol{B})^{-1} + (\boldsymbol{I} - \boldsymbol{Q}^{t+1})(\boldsymbol{I} - \boldsymbol{Q})^{-1}]\boldsymbol{s}. \qquad (14)$$

Because $\boldsymbol{As}$ is the asymptotic response from a single selection, $t\boldsymbol{As}$ is that from t selections, and is the value predicted from classical theory. The

remaining terms in (14) measure the departure from the assumption of a steady rate of response, and reflect the time taken for genes, and thus improvement, to be passed through the population.

Lag in response. The analysis of this departure can be taken further in programmes run for many time periods. Since

$$\lim_{t\to\infty} \boldsymbol{B}^t = \lim_{t\to\infty} \boldsymbol{Q}^t = 0,$$

the difference between the uniform prediction and that expected approaches

$$\lim_{t\to\infty} (\boldsymbol{R}_{(t)} - t\boldsymbol{A}\boldsymbol{r}) = [(\boldsymbol{I}-\boldsymbol{B})^{-1} - (\boldsymbol{I}-\boldsymbol{Q})^{-1}]\boldsymbol{s}$$

$$= \boldsymbol{D}\boldsymbol{s} \tag{15}$$

say, from (14). The matrix $(\boldsymbol{I}-\boldsymbol{Q})^{-1}$ has the following typical form, using the pig example,

$$(\boldsymbol{I}-\boldsymbol{Q})^{-1} = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{array}\right) \tag{16}$$

but $(\boldsymbol{I}-\boldsymbol{B})^{-1}$ has no noticeably simple form. For the pig example, with $\boldsymbol{P}$ given by (3),

$$\boldsymbol{D} = (\boldsymbol{I}-\boldsymbol{B})^{-1} - (\boldsymbol{I}-\boldsymbol{Q})^{-1} = \left(\begin{array}{cc|ccc} -0{\cdot}074 & 0{\cdot}148 & -0{\cdot}185 & 0{\cdot}037 & 0{\cdot}074 \\ -0{\cdot}296 & -0{\cdot}074 & -0{\cdot}407 & -0{\cdot}185 & -0{\cdot}037 \\ \hline -0{\cdot}074 & 0{\cdot}148 & -0{\cdot}185 & 0{\cdot}037 & 0{\cdot}074 \\ -0{\cdot}296 & -0{\cdot}074 & -0{\cdot}407 & -0{\cdot}185 & -0{\cdot}037 \\ -0{\cdot}518 & -0{\cdot}296 & -0{\cdot}630 & -0{\cdot}407 & -0{\cdot}148 \end{array}\right). \tag{17}$$

Because the responses are given by the vector $\boldsymbol{D}\boldsymbol{s}$, columns of $\boldsymbol{D}$ identify the age-sex-class of the original selected animals, and rows of $\boldsymbol{D}$ the subsequent responses in the different age-sex-classes.

Consider response up to time period 20, after the increment in response is fairly steady in our example (Table 2). The rate of response × generations predicted from selection on young males is $(s_1/2L \times 20)$ g/day in live-weight gain. The correction which has to be applied to compute the response either in young males or young females is $-0{\cdot}074 \times s_1$ where the values $-0{\cdot}074$ are the first and third elements of the first column of $\boldsymbol{D}$. Thus the expected response in young males or females from selection in young males is

$$50\,(0{\cdot}2222 \times 20 - 0{\cdot}074) = 218{\cdot}5 \text{ g/day}$$

as shown in Table 2. With selection on both sexes, the correction to be applied to the uniform rate for response in young animals is $-0{\cdot}074 \times 50 - 0{\cdot}185 \times 35 = -10{\cdot}2$ g/day (Table 2).

As equations (14) and (15) show, the difference between $\boldsymbol{R}_{(t)}$ and the prediction based on the uniform response, $t\boldsymbol{A}\boldsymbol{s}$, approaches $\boldsymbol{D}\boldsymbol{s}$ asymptotically. In Table 2 the differences are seen to fluctuate considerably around their

final value of 10·2 g/day in the first few time periods, the prediction becoming useful by about time unit 5 in this example. If older animals were retained for breeding, it would take longer to become a satisfactory prediction.

From (15) and (17) the correction to be applied for response in males or females of age 2 time units is $0{\cdot}222s_1$ or $0{\cdot}222s_3$ greater than for animals of age 1 unit, when selection is practised on young males or females. This is just the rate of response per time period and reflects the fact that the older individuals have been influenced by one time period less selection. There is the same difference, but of opposite sign, between elements of columns 1 and 2 and elements of columns 3 and 4 of $\boldsymbol{D}$. This is because animals do not reproduce in this example until 2 time units of age, so responses from selection just prior to this are realized 1 time unit earlier than if selection is made on the young animals. Since females of age 3 units have already contributed some of their genes, the same simple relationships do not hold between columns 4 and 5. The difference between elements of column 5, 0·111 in (17), reflects the fact that the reproductive value of these oldest breeders is one-half that of individuals of age 1 or 2.

It is illuminating to express the elements of (17) in terms of time rather than genetic improvement. A difference of $1/2L$ between two elements of $\boldsymbol{D}$ is proportional to the response from one time unit, so that

$$\boldsymbol{C} = 2L\boldsymbol{D}$$

with $\boldsymbol{D}$ given by (15), expresses the departure of response from uniform expectation in terms of time periods of selection. This is the lag in the sense used by Bichard (1971). In the pig example $2L = 4{\cdot}5$, and from (17)

$$\boldsymbol{C} = \left(\begin{array}{cc|ccc} -0{\cdot}33 & -0{\cdot}67 & -0{\cdot}83 & 0{\cdot}17 & 0{\cdot}33 \\ -1{\cdot}33 & -0{\cdot}33 & -1{\cdot}83 & -0{\cdot}83 & -0{\cdot}17 \\ \hline -0{\cdot}33 & 0{\cdot}67 & -0{\cdot}83 & 0{\cdot}17 & 0{\cdot}33 \\ -1{\cdot}33 & -0{\cdot}33 & -1{\cdot}83 & -0{\cdot}83 & -0{\cdot}17 \\ -2{\cdot}33 & -1{\cdot}33 & -2{\cdot}83 & -1{\cdot}83 & -0{\cdot}67 \end{array}\right). \tag{18}$$

Letting $\boldsymbol{J}$ be a matrix with all elements to equal to unity, the lag relative to young animals of age 1 unit got by selection in males of age 1 unit can be expressed as

$$\boldsymbol{C} = -\tfrac{1}{3}\boldsymbol{J} + \left(\begin{array}{cc|ccc} 0 & 1 & -\frac{1}{2} & \frac{1}{2} & \frac{2}{3} \\ -1 & 0 & -1\frac{1}{2} & -\frac{1}{2} & \frac{1}{6} \\ \hline 0 & 1 & -\frac{1}{2} & \frac{1}{2} & \frac{2}{3} \\ -1 & 0 & -1\frac{1}{2} & -\frac{1}{2} & \frac{1}{6} \\ -2 & -1 & -2\frac{1}{2} & -1\frac{1}{2} & -\frac{1}{3} \end{array}\right).$$

This equation clearly displays the time lag of the passage of genes. Note for example that there is a lag of one-half of a time unit from selection in females (column 3) rather than males (column 1), since the former are, on average, that much older when their progeny are born.

MULTIPLICATION PROGRAMMES

In the example of a pig population there is only a short time lag in the passage of genes from selected animals to the next generation since animals

are retained for a short time; but it represents a small part of a commercial breeding programme, for there may be several generations of multiplication of stock before commercial animals are produced. Some alternative kinds of multiplication programme have been discussed by Bichard (1971). The issue here is not the efficiency of alternative methods, but solely with showing how a multiplication programme can be fitted into the framework. Again a simple example is used as illustration, by extending the one given previously.

Assume that the pig population defined by (3) is a nucleus herd. After being used in the nucleus, boars are taken to commercial herds and have commercial progeny when 3 time units (1·5 years) of age. Replacement sows are bred in commercial herds, and have progeny when 2, 3 and 4 time units (1, 1·5 and 2 years) of age. For illustration, assume they have $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{6}$ of their progeny at these ages, repectively. There are now three groups of animals to cater for: nucleus males and nucleus females as before, and commercial females (we can exclude commercial males as they have the same breeding value as contemporary commercial females). Thus the new matrix, $\boldsymbol{P}$, can be written

$$\boldsymbol{P} = \left(\begin{array}{ccc|ccc|cccc}
0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{6} & \frac{1}{12} \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}\right) \tag{19}$$

where the blocks of $\boldsymbol{P}$ refer to the passage of genes

$$\left(\begin{array}{c|c|c}
\text{N}\male \text{ to N}\male & \text{N}\female \text{ to N}\male & \text{C}\female \text{ to N}\male \\
\hline
\text{N}\male \text{ to } \textit{N}\female & \textit{N}\female \text{ to N}\female & \text{C}\female \text{ to N}\female \\
\hline
\text{N}\male \text{ to C}\female & \text{N}\female \text{ to C}\female & \text{C}\female \text{ to N}\female
\end{array}\right)$$

and N, C denote nucleus and commercial animals respectively. The blocks corresponding to the passage of genes from commercial to nucleus stock have, of course, all elements equal to zero.

The addition of extra blocks to $\boldsymbol{P}$ does not affect any of the mathematical methods. A row of $\boldsymbol{A}$ for the matrix $\boldsymbol{P}$ given by (19) is

$$\boldsymbol{v}'/2L = (1 \quad 1 \quad 0 \mid 1 \quad 1 \quad 0{\cdot}5 \mid 0 \quad 0 \quad 0 \quad 0)/4{\cdot}5$$

and is the same as that for the nucleus herd with the addition of zero elements for reproductive value outside the nucleus. This merely states that selection outside the nucleus has no long-term effect on improvement, and that the

asymptotic rate of response in any part of the nucleus or multiplication programme is the same, providing all genes in the multiplication herds derive from the nucleus. Of more interest is the difference in mean performance between the nucleus and commercial stock. In intermediate generations equation (14) has to be used; asymptotically the difference depends on the elements of $\boldsymbol{D}$, in terms of genetic progress, or $\boldsymbol{C}$, in terms of time. For this new example,

$$\boldsymbol{C} = \left[\begin{array}{ccc|ccc|cccc} -0{\cdot}33 & 0{\cdot}67 & 0{\cdot}00 & -0{\cdot}83 & 0{\cdot}17 & 0{\cdot}33 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ -1{\cdot}33 & -0{\cdot}33 & 0{\cdot}00 & -1{\cdot}83 & -0{\cdot}83 & -0{\cdot}17 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ -2{\cdot}33 & -1{\cdot}33 & 0{\cdot}00 & -2{\cdot}83 & -1{\cdot}83 & -0{\cdot}67 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ \hline -0{\cdot}33 & 0{\cdot}67 & 0{\cdot}00 & -0{\cdot}83 & 0{\cdot}17 & 0{\cdot}33 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ -1{\cdot}33 & -0{\cdot}33 & 0{\cdot}00 & -1{\cdot}83 & -0{\cdot}83 & -0{\cdot}17 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ -2{\cdot}33 & -1{\cdot}33 & 0{\cdot}00 & -2{\cdot}83 & -1{\cdot}83 & -0{\cdot}67 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ \hline -1{\cdot}50 & -0{\cdot}50 & 4{\cdot}50 & -6{\cdot}50 & -5{\cdot}50 & -2{\cdot}50 & 4{\cdot}50 & 4{\cdot}50 & 2{\cdot}25 & 0{\cdot}75 \\ -2{\cdot}50 & -1{\cdot}50 & 4{\cdot}50 & -7{\cdot}50 & -6{\cdot}50 & -3{\cdot}00 & 4{\cdot}50 & 4{\cdot}50 & 2{\cdot}25 & 0{\cdot}75 \\ -3{\cdot}50 & -2{\cdot}50 & 4{\cdot}50 & -8{\cdot}50 & -7{\cdot}50 & -3{\cdot}50 & 4{\cdot}50 & 4{\cdot}50 & 2{\cdot}25 & 0{\cdot}75 \\ -4{\cdot}50 & -3{\cdot}50 & 4{\cdot}50 & -9{\cdot}50 & -8{\cdot}50 & -4{\cdot}00 & 4{\cdot}50 & 4{\cdot}50 & 2{\cdot}25 & 0{\cdot}75 \end{array}\right] \quad (20)$$

The elements corresponding to genes in the nucleus are, of course, the same as in (18). Consider now the elements in the last block of the first column: these give the asymptotic lag for the passage of genes from males of age 1 time unit in the nucleus to commercial animals. It is necessary to show why there is a difference of $-0{\cdot}33-(-1{\cdot}50) = 1{\cdot}17$ time units between young animals in the nucleus and their commercial contemporaries for improvement deriving from young males, so that the results can be fitted into Bichard's (1971) framework. Consider genes of young nucleus males: 50% come from their sires born 2 time units previously, 25% from their dam's sire born 4·5 time units and 12½% from their dam's grand sire born 7 time units on average previously, giving a weighted value of $0{\cdot}5\times 2{\cdot}0+ 0{\cdot}25\times 4{\cdot}5+ \ldots = 4{\cdot}5$ time units. In the commercial herd, the mean age of females when their progeny are born is 2·67 time units and for males it is 3 time units. The average time taken for genes to pass from young nucleus males to young commercial animals becomes 5·67 units, so the difference between the nucleus and commercial animals of equivalent age is $5{\cdot}67-4{\cdot}5 = 1{\cdot}17$ time units, as (20) shows. The lag is much longer from females bred in the nucleus herd for they are not used for multiplication directly. The values of 4·5 in the third column of $\boldsymbol{C}$ correspond to values of 1·0 in $\boldsymbol{D}$ and merely show that selection among males before passage from nucleus to commercial use realises an ultimate improvement equivalent to half the genetic selection differential.

In Table 2 the predicted responses in the commercial animals of age 1 unit are given for the selection intensities in the nucleus as described previously. The magnitudes of the lag are clearly illustrated, and also the number of time units necessary for this asymptotic lag to become relevant.

Many other examples of multiplication schemes could be given, but most of the principles are illustrated in the previous example. Some extension to the analysis is needed if the commercial animals derive from the cross of the

two populations. Both nucleus populations have to be defined in the matrix P, which affects some of its mathematical properties, but these will not be pursued here.

DIFFERENT SELECTION INTENSITIES FOR BREEDING MALE AND FEMALE REPLACEMENTS

So far it has been assumed that the same selected males and females are used to breed replacements for each sex. Whilst this is the case in many breeding programmes, it is typically not so in dairy cattle AI programmes where only the best animals are used for breeding the limited number of males required for testing. A method for incorporating different selection differentials for breeders of male and female replacements will now be given. This problem has been deferred until now to minimize the initial complexity of the analysis, for some of its nice properties are lost. The fundamental problem is to distinguish between genes of some selected animals which initially pass only to males, for example, but subsequently pass from these males to individuals of both sexes.

Two further matrices E_m and E_f, need to be defined; these specify the passage of genes by reproduction to males only, and to females only, respectively. They are of the same dimension as P, but E_m comprises the first row of P and E_f the first row of the second block of P (i.e. row $h+1$), with all other elements zero. In the pig example (3)

$$E_m = \left(\begin{array}{cc|ccc} 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) \quad \text{and} \quad E_f = \left(\begin{array}{cc|ccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

but it is not a requirement of the analysis that the non-zero rows of E_m and E_f should be the same. Also define s_m and s_f as the vectors of genetic selection differentials of breeders of male and female replacements, respectively.

The vector specifying the response in both sexes from a single cycle of selection is $r_{(t)}$ as previously. In the first time period

$$r_{(1)} = E_m s_m + E_f s_f.$$

The genetic selection differentials applied in the next generation among breeders of males are Qs_m, since the selected group are one time unit older, and similarly for breeders of females. In addition genes may pass from those individuals at time 1 to both sexes. Hence

$$r_{(2)} = Pr_{(1)} + E_m Q s_m + E_f Q s_f. \tag{21}$$

In general

$$r_{(t)} = Pr_{(t-1)} + E_m Q^{t-1} s_m + E_f Q^{t-1} s_f \tag{22}$$

and, of course, the last two terms in (22) vanish when t is sufficiently large that $Q^t = 0$. Alternatively, substituting for $r_{(1)}$ in (21)

$$r_{(2)} = (PE_m + E_m Q)s_m + (PE_f + E_f Q)s_f,$$

and (22) becomes

$$r_{(t)} = \sum_{i=1}^{t} P^{i-1}(E_m Q^{t-i} s_m + E_f Q^{t-i} s_f). \tag{23}$$

Equations (22) and (23) can be used to compute the expected response each generation. It is shown in Appendix 3 that (22) and (23) reduce to the equivalent equation (7) when $s_m = s_f$.

The asymptotic response is, of course, the same in both sexes and (see Appendix 3) can be shown to be

$$\lim_{t\to\infty} r_{(t)} = [(v_m' s_m + v_f' s_f)/2L]\mathbf{1} \tag{24}$$

where v_m and v_f are proportional to the reproductive values of animals only as potential male and female breeders. Thus

$$\left.\begin{aligned} v_{m,i} &= \sum_{j=i}^{h} p_{1j}, \quad v_{f,i} = \sum_{j=i}^{h} p_{h+1,j}, \quad i = 1, \ldots, h \\ v_{m,i} &= \sum_{j=i}^{h+k} p_{1j}, \quad v_{f,i} = \sum_{j=i}^{h+k} p_{h+1,j}, \quad i = h+1, \ldots, k \end{aligned}\right\}. \tag{25}$$

The vector v defined previously is given by

$$v = v_m + v_f,$$

and, if $s_m = s_f$, (24) reduces to (12).

Equation (24) is essentially the well-known formula of Rendel and Robertson (1950) that the asymptotic response is $\Sigma s/\Sigma L$ where Σs and ΣL are the sum of the genetic selection differentials and generation intervals over the four paths of genes, males to males, males to females, females to males and females to females.

The response $R_{(t)}$ from continued selection with the same intensities is obtained most easily from (22) as

$$R_{(t)} = \sum_{T=1}^{t} r_{(T)}.$$

Formulae for the asymptotic value of difference between the total response and that predicted from the uniform rate (24) are derived and given in Appendix 3. They do not have the same simple form as when the same parents are used to breed both sexes.

ECONOMIC EVALUATION OF RESPONSE

A method which is now being commonly used to predict the financial benefits of breeding programmes is to discount returns in future years back to present value (e.g. Poutous and Vissac, 1962; Hinks, 1970; Hill, 1971). Since returns in early generations are discounted least, they can make a large contribution to total discounted returns. It may therefore be important to predict these early responses more accurately than could be done using classical theory for overlapping generations with a uniform rate of response. Incorporation of discounting into the overlapping generation response theory developed here is straightforward.

It is now necessary to include sufficient terms in the matrix P and vectors $r_{(t)}$ and $R_{(t)}$ to enable computations of progress among all animals which

produce commercial returns, not just those responsible for breeding the next generation. These would include all stages of multiplication in pigs, for example, and all ages of lactating cows in a dairy cattle situation. In addition a new row vector $\boldsymbol{w}'$ is required, in which the elements correspond to those of $\boldsymbol{r}_{(t)}$, and are the increment in undiscounted returns (i.e. at current value) from the breeding programme for a unit change in performance. The values comprising $\boldsymbol{w}'$ are assumed to remain constant with time. Let the discount rate be d per time period, and the discount factor be

$$c = 1/(1+d),$$

which is the present value of 1 unit of returns obtained 1 time period from now.

The returns at time t, evaluated at their current value (i.e. at time t) are given by the sum over age-sex groups of the products of responses and returns per unit change. Therefore, the returns from a single year's selection at current value are $\boldsymbol{w}'\boldsymbol{r}_{(t)}$, with $\boldsymbol{r}_{(t)}$ given by (7), where, for simplicity, it is assumed that the same selection differentials are applied to breeders of male and female replacements. If these returns are discounted to present value (i.e. at time 0) the returns, $x_{(t)}$, from a single selection are

$$x_{(t)} = c^t \boldsymbol{w}'\boldsymbol{r}_{(t)}. \tag{26}$$

The total returns, $y_{(T)}$, for the first T time units after the selection is practised, and discounted back to the time of selection are therefore

$$\begin{aligned} y_{(T)} &= \sum_{t=0}^{T} x_{(t)} \\ &= \boldsymbol{w}' \sum_{t=0}^{T} c^t(\boldsymbol{P}^t - \boldsymbol{Q}^t)\boldsymbol{s} \\ &= \boldsymbol{w}'[(\boldsymbol{I}-c\boldsymbol{P})^{-1}(\boldsymbol{I}-c^{T+1}\boldsymbol{P}^{T+1}) - (\boldsymbol{I}-c\boldsymbol{Q})^{-1} \\ &\qquad \times (\boldsymbol{I}-c^{T+1}\boldsymbol{Q}^{T+1})]\boldsymbol{s} \end{aligned} \tag{27}$$

for $c<1$ (i.e. $d>0$), using (7) and (26). If we Take $T\to\infty$ and so discount all future returns, (27) reduces to

$$y_{(\infty)} = \boldsymbol{w}'[(\boldsymbol{I}-c\boldsymbol{P})^{-1} - (\boldsymbol{I}-c\boldsymbol{Q})^{-1}]\boldsymbol{s} \tag{28}$$

which has an appealing simplicity as a solution to an involved problem. If the discount rate is high, the discounted returns after $T = 15$ or 20 years are small, so (28) may not differ much from (27); but even (27) can be computed quickly.

Equation (28) can be expanded as a series (see Appendix 4) which gives some insight into its structure,

$$\begin{aligned} y_{(\infty)} = \boldsymbol{w}'[&\boldsymbol{A}/d + (\boldsymbol{I}-\boldsymbol{B})^{-1} - (\boldsymbol{I}-\boldsymbol{Q})^{-1} - d\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{B})^{-2} + d\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{Q})^{-2} \\ &+ d^2\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{B})^{-3} - d^2\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{Q})^{-3} - \ldots]\boldsymbol{s}, \end{aligned} \tag{29}$$

which converges quickly for commonly used values of $d(<0{\cdot}2$ per year). The first term in (29) is

$$\frac{1}{d}\boldsymbol{A} = \sum_{t=1}^{\infty} t\left(\frac{1}{1+d}\right)^t \boldsymbol{A},$$

which gives the returns predicted from the uniform rate of response. The second pair of terms $(I-B)^{-1}-(I-Q)^{-1}$ specifies the departure due to the asymptotic lag of commercial animals behind those selected, and the remaining smaller terms are additional corrections to these two. Equation (29) can also be used to find the discount rate at which the programme breaks even, with minimum computation, for the matrix inversion and multiplication can be performed first, and then the series computed for a range of scalar multipliers, d.

Consider the pig example including multiplication given by (19), with selection intensities on young animals in the nucleus as before. Assume that the increment in net returns from improving live-weight gain in one pig is £0·01 per 1 g/day (an arbitrary figure). The nucleus comprises 100 sows, and from it 800 male and 700 female bacon pigs are slaughtered per year with the rest retained for breeding and a further 2000 pigs are slaughtered each year in the commercial herd. Thus

$$\boldsymbol{w}' = (8 \quad 0 \quad 0 \mid 7 \quad 0 \quad 0 \mid 20 \quad 0 \quad 0 \quad 0).$$

With $d = 0{\cdot}0488$ per time period (corresponding to an annual discount rate of 10%), equation (28) gives $y_{(\infty)} = £12\,300$.

With continued selection, formulae such as (27) can be extended, but become rather involved. It is probably easier to compute the returns directly as the sum,

$$Y_{(T)} = \sum_{t=0}^{T} c^t y_{(t)}$$

of returns of selection at each time period, $y_{(t)}$ given by (27), discounted back to the beginning. In the limiting case of discounting up to $T \to \infty$,

$$Y_{(\infty)} = y_{(\infty)}/d.$$

Performance of selected animals. The evaluation of response has solely included improvement in the performance of progeny as a result of selection of parents. In species in which traits are expressed several times during the animal's life, for example wool yield or reproductive performance in sheep and milk yield in cattle, selection among young animals for such traits improves subsequent performance in the flock or herd when these animals are retained, providing the repeatability of the trait is not zero. Prediction of this improvement is easily included in the analysis.

The response has been given in terms of the vector of genetic selection differentials, $\boldsymbol{s}$, which is equivalent to phenotypic selection differentials of $\boldsymbol{s}$/heritability and changes in subsequent expressions of the trait of $\boldsymbol{z} = \boldsymbol{s} \times$ repeatability/heritability. Assuming annual breeding and expression of the trait, the improvement among selected animals is therefore $\boldsymbol{Qz}$, $\boldsymbol{Q}^2\boldsymbol{z}$ in successive years. Therefore the total discounted improvement in mean performance of the population t years after selection is, using (26),

$$x_{(t)} = c^t \boldsymbol{w}'[(\boldsymbol{P}^t - \boldsymbol{Q}^t)\boldsymbol{s} + \boldsymbol{Q}^t\boldsymbol{z}].$$

DISCUSSION

In a population in which generations are discrete the pattern of response can be described and predicted adequately by just the mean performance of

the group of animals born in the current generation (or time period since this can be taken as equal to the generation interval); but with overlapping generations it is necessary to describe the performance of all age groups present in the population at any time and so a scalar description of the population has been replaced by a vector description. Predictions of response to selection then involve matrix rather than scalar algebra, so the theory presented here is a natural extension of that for non-overlapping generations. Whilst all the results could be obtained without formal use of matrix algebra, it considerably simplifies the presentation and lends itself to computer operations.

The predictions of classical overlapping generation theory hold only asymptotically, since with recurrent selection of constant intensity, a fixed relationship between the genetic improvements in the different age groups is not established immediately. The main objective of the analysis has been to facilitate predictions of response before the asymptotic state has been reached, so although the paper is primarily a presentation of methodology, some of the features of the irregularity of response have been illustrated in the simple examples of pig populations. There are much more pronounced departures from uniform responses in dairy cattle with AI testing programmes, but these would have required definition of larger matrices involving more space and, perhaps, exhibiting the major points less clearly.

Some of the more important assumptions made in the analysis should be emphasized. The population structure, specifically the parental age distribution, is assumed to remain constant (but it is not always appreciated that this assumption is made in the classical theory of selection with overlapping generations). To avoid this assumption, $\boldsymbol{P}$ could be replaced by a time-dependent matrix $\boldsymbol{P}_{(t)}$, giving $\boldsymbol{r}_{(t)} = \boldsymbol{P}_{(t)}\boldsymbol{r}_{(t-1)}$, but unless changes in $\boldsymbol{P}_{(t)}$ were known *a priori* such a theory would have little predictive value. It is implicitly required that the population not be very small, for the parental age distribution would then deviate by chance from that expected. We have assumed that the genetic selection differentials remain constant with repeated selection, implying that the parameters such as heritability and variance do not change. It is also assumed that there is little departure from additive gene action, for no terms for inbreeding depression have been included. All these assumptions will be less tenable if the population is of small effective size.

Computation of the genetic selection differential is not as straightforward in the overlapping generation model as is often assumed. Bichard *et al.* (1973) have pointed out that the younger parents in the population are the result of more years of selection, and thus are expected to have a higher breeding value. So if the optimum selection scheme is being practised the parental age distribution may depart considerably from that based on the assumption of genetically homogeneous parental age groups.

Rather uncritical use has been made here of the discounting procedure for computing monetary returns, yet many important assumptions have to be made outwith genetics, such as the size and constancy of the market for breeding stock, and the value of improvement of individual animals (Hill, 1971). One may question whether sophistications of prediction of response with overlapping generations are therefore necessary when computing discounted returns. In some cases a uniform prediction of response may be adequate (Hill, 1971), but as C. J. M. Hinks (personal communication) has

pointed out, this is more likely to be so when comparisons are being made of alternative schemes which merely affect the magnitudes of rates of response and not the timing of it. If alternative uses of capital, such as for performance or progeny testing, are being compared and these involve very different patterns of response, then the timing of the improvement is more critical. Exact predictions such as can be made using the theory developed here then seem appropriate, and the methods suggested are straightforward.

ACKNOWLEDGEMENTS

I am indebted to Professor Alan Robertson and Dr Brian McGuirk for several helpful comments and suggestions and to Mrs Kathy Burgoyne for computing the examples.

REFERENCES

BICHARD, M. 1971. Dissemination of genetic improvement through a livestock industry. *Anim. Prod.* **13**: 401–411.

BICHARD, M., PEASE, A. H. R., SWALES, P. H. and ÖZKÜTÜK, K. 1973. Selection in a population with overlapping generations. *Anim. Prod.* **17**: 215–227.

DICKERSON, G. E. and HAZEL, L. N. 1944. Effectiveness of selection on progeny performance as a supplement to earlier culling in livestock. *J. agr. Res.* **69**: 459–476.

HILL, W. G. 1971. Investment appraisal for national breeding programmes. *Anim. Prod.* **13**: 37–50.

HILL, W. G. 1972. Effective size of populations with overlapping generations. *Theor. Pop. Biol.* **3**: 278–289.

HINKS, C. J. M. 1970. The selection of dairy bulls for A.I. *Anim. Prod.* **12**: 569–576.

HINKS, C. J. M. 1971. The genetic and financial consequences of selection amongst dairy bulls in artificial insemination. *Anim. Prod.* **13**: 209–218.

HINKS, C. J. M. 1972. The effects of continuous sire selection on the structure and age composition of dairy cattle populations. *Anim. Prod.* **15**: 103–110.

KEMENY, J. G. and SNELL, J. L. 1960. *Finite Markov Chains.* van Nostrand, Princeton, New Jersey.

LESLIE, P. H. 1945. On the use of matrices in certain population mathematics. *Biometrika* **33**: 183–212.

POUTOUS, M. and VISSAC, B. 1962. Recherche théoretique des conditions de rentabilité maximum de l'épreuve de descendance des tauraux d'insemination artificielle. *Annls Zootech.* **11**: 233–256.

RENDEL, J. M. and ROBERTSON, A. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. *J. Genet.* **50**: 1–8.

SEARLE, S. R. 1961. Estimating herd improvement from selection programmes. *J. Dairy Sci.* **44**: 1103–1112.

VAN VLECK, L. D. 1964. Sampling the young sire in artificial insemination. *J. Dairy Sci.* **47**: 441–446.

(*Received* 2 *August* 1973)

APPENDIX

1. *Computation of A*

Since the elements of each row of $\boldsymbol{P}$ are non-negative and sum to unity, it is a stochastic matrix and the relevant theory for such matrices (e.g. Kemeny and Snell, 1960) can be used. With the elements of $\boldsymbol{P}$ specified, the stochastic matrix has only one ergodic state, so $\boldsymbol{P}$ has a single eigenvalue of unity and all others are of smaller absolute value. Thus the matrix $\boldsymbol{A}$, given by $\boldsymbol{A} = \lim_{t \to \infty} \boldsymbol{P}^t$, is associated with the unit eigenvalue. Since the row sums of $\boldsymbol{P}$ are all the same, the vector $\mathbf{1}$ (with all elements equal to unity) is

a right eigenvector of P. We also find that v' (given by (10)) is a left eigenvector of P. For example the ith element of $v'P$, for $1 \leqq i \leqq h-1$, is given by

$$v_1 p_{1i} + v_{h+1} p_{h+1,i} + v_{i+1} = p_{1i} + p_{h+1,i} + \sum_{j=i+1}^{h} (p_{1j} + p_{h+1,j})$$

$$= v_i,$$

and this result holds for other values of i. Since v' and $\mathbf{1}$ are eigenvectors, $A = \alpha \mathbf{1} v'$ for some constant α, and, because P is stochastic, A is stochastic; hence a row total of A satisfies $\alpha \sum_i v_i = 1$. From (10),

$$\alpha^{-1} = \sum_{i=1}^{h+k} v_i$$

$$= \sum_{i=1}^{h} \sum_{j=i}^{h} (p_{1j} + p_{h+1,j}) + \sum_{i=h+1}^{h+k} \sum_{j=i}^{h+k} (p_{1j} + p_{h+1,j})$$

$$= \sum_{i=1}^{h} i(p_{1i} + p_{h+1,i}) + \sum_{i=1}^{k} i(p_{1,h+1} + p_{h+1,h+i})$$

$$= 2L.$$

Thus $A = \mathbf{1}v'/2L$ as given in (11), and

$$\lim_{t \to \infty} r_{(t)} = As$$

$$= \mathbf{1}(v's)/2L = (v's/2L)\mathbf{1}$$

as given in (12), since $v's$ is a scalar.

2. *Geometric series in P and Q*

The sum, from (13), of $\sum_{i=0}^{t} P^t$ is required but since P has an eigenvalue of unity, the series does not converge as $t \to \infty$. If P is partitioned as $P = A + B$ then, since P is the first term in the spectral decomposition of P, $AB = BA = 0$; and $A^t = A$, $t \geqq 1$. Hence

$$P^t = A + B^t, \; t \geqq 1.$$

Therefore

$$\sum_{i=0}^{t} P^t = tA + \sum_{i=0}^{t} B^t = tA + (I - B^{t+1})(I - B)^{-1}$$

since all eigenvalues, λ, of B satisfy $|\lambda| < 1$, because P has only one unit eigenvalue.

The matrix Q is triangular with all diagonal elements, and therefore all eigenvalues, equal to zero; so

$$\sum_{i=0}^{t} Q^t = (I - Q^{t+1})(I - Q)^{-1}$$

giving (14).

3. *Different selection intensities for breeders of each sex*

Reduction of (23) *to* (7) *when* $s_m = s_f$. With equality of selection differentials (23) reduces to

$$r_{(t)} = \sum_{i=1}^{t} P^{i-1}(E_m + E_f)Q^{t-i}s.$$

But $P = E_m + E_f + Q$, so

$$\begin{aligned} r_{(t)} &= \sum_{i=1}^{t} (P^iQ^{t-i} - P^{i-1}Q^{t-i+1})s \\ &= (P^t - Q^t)s \end{aligned}$$

as in (7), since all other terms cancel.

Proof of (24). Equation (23) may be rewritten, using $P = A + B$, as

$$r_{(t)} = \left[\sum_{i=2}^{t} AE_mQ^{t-i} + \sum_{i=1}^{t} B^{i-1}E_mQ^{t-i} \right] s_m \tag{1A}$$

plus terms in E_f, s_f which are ignored for the present. Now for $t \geqq \max(h, k) = h$, say, $Q^t = \mathbf{0}$. Thus

$$\begin{aligned} \lim_{t\to\infty} r_{(t)} &= \left[AE_m(I-Q)^{-1} + \lim_{t\to\infty} \sum_{i=0}^{h-1} B^{t-i}E_mQ^i \right] s_m \\ &= AE_m(I-Q)^{-1}s_m \end{aligned}$$

since $\lim_{t\to\infty} B^t = \mathbf{0}$. The elements of $(I-Q)^{-1}$ are illustrated in (16) and it is found that

$$E_m(I-Q)^{-1} = \begin{pmatrix} v'_m \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

where v'_m is given by (25). The first element of each column of A is $v/2L = 1/2L$. Hence

$$AE_m(I-Q)^{-1}s_m = \frac{1}{2L}\begin{pmatrix} v'_m \\ v'_m \\ \vdots \\ v'_m \end{pmatrix} s_m = (v'_m s_m/2L)\mathbf{1}. \tag{2A}$$

Adding the similar term for breeders of females gives (24),

$$\lim_{t\to\infty} r_{(t)} = [(v'_m s_m + v'_f s_f)/2L]\mathbf{1}.$$

Departure from uniform rate of response. Whilst a general result would be useful the vector

$$\lim_{t\to\infty} \left[R_{(t)} - \frac{t}{2L}(v'_m s_m + v'_f s_f)\mathbf{1} \right]$$

is specifically required, for no simple closed form for the value of this difference at all values of t has been obtained.

From (1A), again ignoring terms in E_f, s_f for the present,

$$\boldsymbol{R}_{(t)} = \sum_{T=1}^{t} \boldsymbol{r}_{(t)}$$

$$= \left[A\boldsymbol{E}_m(\boldsymbol{I}-\boldsymbol{Q})^{-1} \sum_{T=1}^{t} (\boldsymbol{I}-\boldsymbol{Q}^{T-1}) + \sum_{T=1}^{t} \sum_{i=1}^{T} \boldsymbol{B}^{i-1}\boldsymbol{E}_m\boldsymbol{Q}^{T-i} \right] s_m$$

giving

$$\lim_{t\to\infty} [\boldsymbol{R}_{(t)} - tA\boldsymbol{E}_m(\boldsymbol{I}-\boldsymbol{Q})^{-1}s_m] = [-A\boldsymbol{E}_m(\boldsymbol{I}-\boldsymbol{Q})^{-2}$$

$$+(\boldsymbol{I}+\boldsymbol{B}+\boldsymbol{B}^2+\ldots)\boldsymbol{E}_m(\boldsymbol{I}+\boldsymbol{Q}+\ldots+\boldsymbol{Q}^h)]s_m.$$

Using (2A), and including terms in s_f

$$\lim_{t\to\infty}\left[\boldsymbol{R}_{(t)} - \frac{t}{2L}(v'_m s_m + v'_f s_f)\mathbf{1}\right] = [(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{E}_m - A\boldsymbol{E}_m(\boldsymbol{I}-\boldsymbol{Q})^{-1}](\boldsymbol{I}-\boldsymbol{Q})^{-1}s_m$$

$$+[(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{E}_f - A\boldsymbol{E}_f(\boldsymbol{I}-\boldsymbol{Q})^{-1}](\boldsymbol{I}-\boldsymbol{Q})^{-1}s_f \qquad (3A)$$

$$= \boldsymbol{D}_m s_m + \boldsymbol{D}_f s_f.$$

The two matrices $\boldsymbol{D}_m$ and $\boldsymbol{D}_f$ define the departure from the uniform response prediction, and $\boldsymbol{D}_m/2L$, $\boldsymbol{D}_f/2L$ are the matrices defining this departure in terms of the time lag.

Using the relations $\boldsymbol{P} = \boldsymbol{Q}+\boldsymbol{E}_m+\boldsymbol{E}_f = A+\boldsymbol{B}$ in (3A)

$$\boldsymbol{D}_m+\boldsymbol{D}_f = (\boldsymbol{I}-\boldsymbol{B})^{-1}[A-(\boldsymbol{I}-\boldsymbol{B})+(\boldsymbol{I}-\boldsymbol{Q})](\boldsymbol{I}-\boldsymbol{Q})^{-1}$$

$$-A[A-\boldsymbol{I}+\boldsymbol{B}+(\boldsymbol{I}-\boldsymbol{Q})](\boldsymbol{I}-\boldsymbol{Q})^{-2}. \qquad (4A)$$

Noting that $(\boldsymbol{I}-\boldsymbol{B})^{-1}A = A$ since $\boldsymbol{B}A = \mathbf{0}$ and $A^2 = A$, (4A) reduces to

$$\boldsymbol{D}_m+\boldsymbol{D}_f = (\boldsymbol{I}-\boldsymbol{B})^{-1}-(\boldsymbol{I}-\boldsymbol{Q})^{-1} = \boldsymbol{D}$$

agreeing with (14).

4. *Series relating to discounting*

For $d>0$, the matrix $\frac{1}{1+d}\boldsymbol{P}$ has all eigenvalues, λ, satisfying $|\lambda|<1$, so

$$\sum_{t=0}^{\infty}\left(\frac{1}{1+d}\boldsymbol{P}\right)^t = \left(\boldsymbol{I}-\frac{1}{1+d}\boldsymbol{P}\right)^{-1}. \qquad (5A)$$

Equation (5A) can be expanded as follows

$$\sum_{t=0}^{\infty}\left(\frac{1}{1+d}\boldsymbol{P}\right)^t$$

$$= \sum_{t=1}^{\infty}\left(\frac{1}{1+d}A\right)^t + \sum_{t=0}^{\infty}\left(\frac{1}{1+d}\boldsymbol{B}\right)^t$$

$$= \frac{1}{d}A + \sum_{t=0}^{\infty}[(1-d+d^2-\ldots)\boldsymbol{B}]^t$$

$$= \frac{1}{d}\boldsymbol{A} + \sum_{t=0}^{\infty} \boldsymbol{B}^t - d\boldsymbol{B} \sum_{t=0}^{\infty} (t+1)\boldsymbol{B}^t + d^2\boldsymbol{B} \sum_{t=0}^{\infty} \tfrac{1}{2}(t+1)(t+2)\boldsymbol{B}^t - \ldots$$

$$= \frac{1}{d}\boldsymbol{A} + (\boldsymbol{I}-\boldsymbol{B})^{-1} - d\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{B})^{-2} + d^2\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{B})^{-3} - \ldots$$

Similarly,

$$\sum_{t=0}^{\infty} \left(\frac{1}{1+d}\boldsymbol{Q}\right)^t = (\boldsymbol{I}-\boldsymbol{Q})^{-1} - d\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{Q})^{-2} + d^2\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{Q})^{-3} - \ldots$$

18

Theoretical aspects of crossbreeding

by

William G. Hill

Ann. Génét. Sél. anim., 1971, **3** (1), 23-34.

THEORETICAL ASPECTS OF CROSSBREEDING

W. G. HILL

Institute of Animal Genetics, Edinburgh EH9 3JN, Scotland

Annales de Génétique et de Sélection animale
INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE
149, rue de Grenelle, Paris-7e

Ann. Génét. Sél. anim., 1971, **3** (1), 23-34.

THEORETICAL ASPECTS OF CROSSBREEDING*

W. G. HILL

Institute of Animal Genetics, Edinburgh EH9 3JN, Scotland

SUMMARY

Methods of utilising breeds and breed crosses in animal production are discussed, taking account of both genetical and economic aspects. The theoretical principles for breed and breed cross comparison are analysed, but most emphasis is given to methods of improvement of existing crosses. A new synthetic breed is likely to have higher genetic variation, and reach a higher selection limit than the pure breeds from which it originates. However, it may take many years for the synthetic to surpass the best available purebred under continuous selection. Returns obtained in early years have more monetary benefit than those obtained later, for they can earn interest and incur a smaller risk element, so that a synthetic of use only in later years is unlikely to be cost-effective. Despite the flexibility in maintaining several alternative breeds, these need to be continually selected if they are to remain competitive, so better returns may be obtained by exerting more pressure on the best available present material. It is unlikely on theoretical grounds that cross testing schemes such as reciprocal recurrent selection have much to offer for breed cross improvement in large animals where growth and carcase traits are important.

INTRODUCTION

Crossbreeding has been an established practice for centuries in the domesticated animal species. Breeders have had many objectives: crosses have been made every generation to obtain any benefits there may be from heterosis or from the particular merits of the individual breeds as maternal or paternal parents. Alternatively the crosses have been used to form new populations with desirable characters from each of the parental breeds with, perhaps, increased variability to enable more rapid progress from later selection. The theoretical basis of crossbreeding has been studied extensively to enable us both to understand the genetic mechanism underlying heterosis and to design breeding programmes to utilise it.

(*) Invited report presented in the Study Meeting of the European Association for Animal Production, Genetic Commission, Gödöllö, Hungary, august 24 th, 1970.

There are two essentially separate aspects of crossbreeding, although they can not be considered entirely independently of each other. The first includes the choice of breeds and method of utilising them in crosses, if necessary, in order to maximise *present* economic performance. For example, we may wish to know whether breed cross A × B is superior to A × C or to A as a single breed, when all productive and maternal traits are considered. The second area of breed utilisation is concerned with improvement over a period of a few generations. We would like to know which breeds or crosses to choose now and use in a selection programme so as to maximise economic merit over the next 10 or 20 years. The extreme examples occur with corn or poultry breeding, using a cross of inbred lines. The breeder may have the best two-way cross on the market at present, but could find difficulty improving it. There is some suggestion that breeding programmes in corn are moving back from an inbreeding and crossing scheme towards programmes in which selection is practised every generation. In the large animal context we are more concerned with whether to form new breeds with, perhaps, enhanced variation, or whether to use the best available at present.

In a recent review DICKERSON (1969) discussed the experimental information required for a rational choice of breeds, but was primarily concerned with immediate performance. Although I shall briefly discuss the theoretical framework on which such decisions should be made, I will give more emphasis to the problem of maximisation of future performance which has not, I believe, been investigated adequately in the context of breed utilisation. Unfortunately the analysis is bound to be somewhat speculative, for we generally lack adequate information on genetic parameters within different breeds and crosses in most practical situations. However it is possible to set out some of the conditions under which new cross populations might respond faster and further than their parent breeds. The analysis has not been taken very far, but hopefully it will provide a few pointers, and I shall give more attention to the arguments on which decisions should be based, rather than to conclusions in any specific instance.

For the purpose of this discussion the term *breed* will refer to any closed population from which members can be identified by phenotype or pedigree. A breed may have been kept distinct from other breeds under consideration for only a few generations, so that, for example, Canadian and Dutch Holstein cattle may be viewed as separate breeds for this purpose. I shall also make considerable reference to *productive* and *maternal* traits. In the class of productive traits are included growth and carcase characters of animals for slaughter for meat and milk production in a dairy breed. Maternal traits include litter number, conception rates, milk production in suckler herds and perhaps even adult body size, in so far as it affects breeding costs. In effect, the genes for productive traits are contributed by both parents in a cross, those for maternal traits are expressed only in the dam. The other term to be defined is *synthetic*, which will be used for any new breed cross which is maintained as a new population, breed or " gene pool ".

CROSSBREEDING AND PRESENT PERFORMANCE

In principle, the utilisation of crossbreds to obtain maximum performance at the present is simple. It is necessary only to find the most efficient purebred or crossbred combination, taking account of both productive and maternal traits. There may, however, be considerable difficulties in actually finding the best cross combination, especially when there are specific heterotic relationships between pairs of breeds and when there are important genotype by environment interactions. In these situations it may be necessary to test a large number of combinations. Otherwise good predictions of merit may be possible from pure line performance in some standard environment. MOAV (1966) discussed criteria for evaluating crossbreds. He defines a non-linear relation between maternal performance and economic merit, but we shall simplify this here to linearity. Consider a cross of breeds A (sire) and B (dam) with productive performance P_A, P_B and heterosis P_{AB}, and for the dam breed a maternal performance R_B. The economic merit, E, is

$$E = K + x\left(\frac{1}{2}P_A + \frac{1}{2}P_B + P_{AB}\right) + y\,R_B$$

or in a three-way cross A × (B × C) it is, approximately,

$$E = K + x\left(\frac{1}{2}P_A + \frac{1}{4}P_B + \frac{1}{4}P_C + \frac{1}{2}P_{AC} + \frac{1}{2}P_{BC}\right) + y\left(\frac{1}{2}R_B + \frac{1}{2}R_C + R_{BC}\right)$$

Here K, x and y are appropriate constants. Of these K includes fixed costs and does not affect comparisons between breeds. Examples of the values of x and y are given by MOAV (1966) for pigs, and these can be modified to correspond with the formulation used here. Let E be the excess of returns over variable costs, measured in pounds sterling per pig of 100 kg live weight marketed. Letting P be the feed conversion efficiency (kg feed per kg gain) then $x = 3.1$, and letting R be the number of pigs marketed per sow per year then $y = 0.21$, where R has a mean of about 16. These figures are for integrated operations, and they may not reflect present economic conditions, but should serve as an example.

These formulae illustrate some important, if somewhat obvious, points. Unless there is a large amount of interaction, P_{AB}, *specific* to particular breed combinations, the sire breed with highest performance on productive traits should be used, for we are assuming here that many dams are mated per sire, or that AI is used, so that the sire breed contributes a very small proportion of total maintenance costs. In the dam breed both productive and maternal traits have to be considered, and the weightings x and y determine how much should be given to each. These same weightings can be used for calculating indices for selection within breeds. We see that the fixed crossing scheme takes full advantage of any heterosis for productive traits in a two-way cross, and for maternal traits also in the three-way cross.

In cattle or sheep a high proportion of animals may have to be bred pure to provide replacements in the dam breed. If a proportion, q, of the animals marketed are pure breds of the dam breed, and $1-q$ are crosses, the average merit becomes

$$E = K + x\left[\left(\frac{1-q}{2}\right)P_A + \left(\frac{1+q}{2}\right)P_B + (1-q)P_{AB}\right] + yR_B$$

so that productive performance in the dam breed becomes relatively more important. If a new synthetic breed is made from the cross of the A and B breeds the overall merit becomes

$$E = K + x\left(\frac{1}{2}P_A + \frac{1}{2}P_B + \frac{1}{2}P_{AB}\right) + y\left(\frac{1}{2}R_A + \frac{1}{2}R_B + \frac{1}{2}R_{AB}\right)$$

There is a loss of half the heterozygosity for productive traits, but a gain in the maternal traits. With a rotational crossing scheme on two breeds the average merit, taken over successive crosses, includes $2/3$ of the heterosis between the breeds for productive and maternal traits, but is otherwise the same as for the synthetic.

This discussion will not be carried further here. Reference should be made to the papers of DICKERSON (1969), MOAV (1966) and FEWSON and JAKUBEC (1970).

CROSSBREEDING AND FUTURE PERFORMANCE

In making decisions about breed or breed cross improvement in future years we face problems at two levels. We have to estimate the potential genetic progress and compare these rates of progress with alternative schemes. In addition we should consider the costs of these schemes and relate these to their potential economic benefits. Most geneticists have occupied themselves with measurement of response, considering economics only when designing a selection index to give optimum weight to the traits. I feel we need to go further than this and will attempt to do so after some discussion of the relevant genetic theory.

Imagine that on the basis of our breed and breed cross testing programme we find that the breed cross $A_1 \times B_1$ is most efficient. Therefore, unless there are specific interactions between these breeds, A_1 is the best available for productive traits and B_1 is good for both productive and maternal characters. We now have several options open for improving the cross, although some of them may not seem very promising. These are: (*a*) form a synthetic from the $A_1 \times B_1$ cross; (*b*) select solely within the breeds A_1 and B_1; (*c*) initiate rotational crossbreeding between A_1 and B_1; (*d*) form a synthetic sire or dam breed; and (*e*) maintain alternative sire or dam lines. The options are not mutually exclusive not do they cover the whole range of possible programmes, but they give some indication of the main direction of selection effort. We shall consider them in turn.

A. — *Form synthetic from* $A_1 \times B_1$ *cross.*

A new breed could be formed and maintained and marketed as a pure breed but this is unlikely to be useful. There is an initial loss of half the heterosis between the breeds for productive traits, which later increases as the synthetic becomes inbred, and a loss of half the maternal advantage of breed B_1 over A_1. Secondly, it has been shown by SMITH (1964) and MOAV and HILL (1966) that greater progress for overall merit is made if separate sire and dam lines are maintained, with selection in the sire line (or breed) made solely for productive traits and in the dam line for an index of productive and maternal traits. This advantage may be small in species such as pigs in which important maternal traits all have low heritability so that little pressure should be imposed on them.

In a dual purpose beef and dairy cattle system there may be considerable advantages in maintaining separate breeds. In the dam, or dairy breed, most selection effort has to be applied to milk production, and selection for beef characteristics can only be undertaken with minor weighting in the milk progeny test, or by performance testing prior to the progeny test. In either case the rate of response for traits relevant to beef production is much smaller than could be achieved in a beef breed used solely as a sire in crosses. In the beef breed intense selection can be practised on a performance test, using a short generation interval. Imagine, for example, that a pure Holstein could currently outperform any cross with the Holstein on some intensive management systems. Yet after a few years of selection either in a beef breed or in a separate strain of Holsteins, crosses to this breed or strain could be superior for beef traits, so that cross matings in excess of requirements for dairy breed replacement should be made.

There may be an increase in variability in the $A_1 \times B_1$ cross relative to the parent lines so that response is enhanced. However there are more appropriate means of forming synthetics with the aim of increasing variation, and these are discussed later.

B. — *Select within* A_1 *and* B_1 *breeds.*

In this way we retain, at least in the short term, the heterosis and other desirable properties of the cross combination. The main issue in this scheme is the mode by which selection should be practised: whether it should be based on pure line or on cross performance using some scheme such as reciprocal recurrent selection. For traits determined primarily by additive or completely dominant genes it has been shown theoretically that the rate of improvement in the cross and the selection limit are approximately the same in pure line and reciprocal recurrent selection schemes, *providing* that the same intensity of selection is practised in each system (HILL, 1970). But it is unlikely that any improvement scheme using cross testing could be operated in large animals with the same intensity and generation interval as in schemes for within breed selection, except perhaps

in programmes to improve milk production using progeny testing. If there is overdominance faster rates and higher limits can, of course, be achieved with reciprocal recurrent selection. An indication of whether this might be possible can be obtained from the genetic correlation of pure and cross performance. If this is close to unity there will be no advantage in the short term in selecting for cross performance directly. However, it is conceivable, in theory at least, that an initial programme of pure line selection would reduce later gains with reciprocal recurrent selection when both breeds have approximately the same gene frequency so that there is selection towards the equilibrium frequency. In large animals the traits of major importance include growth rate (and feed conversion efficiency), carcase quality (or simply degree of fatness), milk yield and milk quality, and reproductive traits. Of these carcase and milk quality typically show little heterosis, growth rate and milk production moderate heterosis, and the reproductive traits exhibit rather more. One can conjecture therefore that at most only a small proportion of the variance for all these traits, with the possible exception of fertility and litter size, for example, are contributed by overdominant genes. Breeding programmes with selection on pure line performance can therefore be continued with safety.

Whilst there appears to be little place for selection programmes based on cross performance in a two way cross structure they could be more relevant for improving the reproductive performance of the B × C mother in the three-way cross A × (B × C). But although each breed in the dam side of the cross contributes only 1/4 of the genes for the productive traits in the final crossbred animals it also contributes only 1/2 to the maternal performance of B × C. The relative index weightings which should be applied to maternal and productive traits in these breeds B and C are therefore almost the same as should be used in the single dam breed of a two way cross. In pigs the economic weightings for food conversion efficiency and carcase quality are so high, and the heritability of litter size is so low that most selection pressure should be devoted to these productive traits in the dam breeds. Thus even in a three-way cross a reciprocal recurrent selection programme would seem unjustified. Similarly, inbreeding schemes used to generate between line variation within the chosen breeds can not be effective relative to programmes utilising constant selection for the highly heritable traits.

C. — *Rotational crossbreeding of A_1 and B_1.*

In a rotational crossbreeding scheme each breed contributes to the cross to the same extent on average, both as a sire breed and as part of the dam combination. Therefore selection pressure has to be put on the same traits, both productive and maternal, in each of the two (or more) parent breeds, so that specialised sire and dam lines can not be developed. We must then expect to make less selection progress in the rotational crossbred than in a fixed crossing scheme such as $A_1 \times B_1$, where different programmes can be used for the two breeds.

D. — *Form synthetic sire or dam breed.*

If we have available other breeds A_2, A_3 etc. which are only slightly poorer than A_1 as sire breeds, these could be crossed with A_1 to form a synthetic and yet retain general heterosis in the cross. Similarly other dam breeds B_2, B_3 could be crossed with B_1 to form a synthetic sire line. These are likely to be more attractive alternatives than making a synthetic from the cross $A_1 \times B_1$. The new synthetic breeds could be useful if they show greater genetic variation thanthe pure breeds, so that after a few years of selection their merit will reach and then surpass that of A_1 or B_1, and could then be substituted in the cross. JAMES (1966) has discussed procedures for selecting animals from among several populations, but only in the context of maximising the present performance of the synthetic.

If there is information available on heritabilities in the breed A_1 and the synthetic A_{1x2}, say, it is simple to predict the time needed before it surpasses A_1. However this could be many years in a practical situation. For example, assume that in beef breeds the trait, weight to 400 days, has a standard deviation of 40 kg and that A_1 exceeds A_{1x2} by 10 kg (in breeding value since heterosis within the sire line is not of interest). In an efficient breeding programme with selection only on males and rapid replacement of females an annual response of 16 h^2 kg per year can be made. So if the heritability in the synthetic was, say, 50 % and in the pure breed it was 40 % and both were continuously selected, it would take $10/(16 \times 0.1)$ or at least six years for the new breed to catch up. Some years would also be needed to establish and multiply the synthetic and obtain the necessary estimates of genetic parameters.

It is usually difficult or expensive to obtain accurate estimates of heritability, and it is unlikely in many situations that estimates of *differences* in heritability between synthetics and pure breeds could be obtained with sufficient precision that practical decisions could be taken using them. It is possible to make some theoretical predictions of differences in genetic variance, but these too suffer from severe limitations. The simplest situation is where breeds A_1 and A_2, say, are essentially randomly selected but distant by several generations from a common base. Assume there is additive gene action, and the additive variance in the synthetic (or in the foundation population) is σ^2_0. If the populations have been inbred by an amount F, the expected within-population variance is $(1 - F)\sigma^2_0$, and the variance between populations is $2F\sigma^2_0$. In a sample of size two from a normal distribution the first ranking individual is, on average, 0.56 standard deviations superior to the mean of the two. If h^2 is the heritability in the foundation or in the synthetic population, and the phenotypic variance is assumed to be altered, the synthetic will take about $0.56\sqrt{2F}/iFh$ generations to reach the better pure line when both are under continued selection. For example, if $F = 0.2$, $i = 1.0$ (averaged over sexes) and $h^2 = 0.4$, the synthetic is expected to take 2.8 generations to reach the better pure line, or 7 years for our beef cattle example with the 2.5 year generation interval. After that period, assuming

there had been no change in variance through selection or further inbreeding, the synthetic would gradually become increasingly superior.

In other cases predictions of variance in the synthetic are essentially speculative, although one or two useful relationships are known. Let q_1 and q_2 be the frequency of some gene in lines A_1 and A_2, and $\bar{q}$ be the mean frequency. Then

$$\bar{q}(1-\bar{q}) = \frac{1}{2}q_1(1-q_1) + \frac{1}{2}q_2(1-q_2) + \frac{1}{4}(q_1-q_2)^2$$

so the mean heterozygosity at this locus and variance if the genes act additively is at least as high in the synthetic as in the average of the two parental lines. More generally, JACKSON and JAMES (1970) have shown that, with additive effects, the variance within the synthetic is given by $\frac{1}{2}\sigma^2_B + \sigma^2_W$, where σ^2_B is the genetic variance between populations and σ^2_W the genetic variance within populations, assumed to be the same in each. At loci showing complete dominance the additive variance is higher in the synthetic when the mean frequency of the recessive allele is greater than 0.5, otherwise it is less (LERNER, 1954). But at such loci most additive variance is expressed when the recessive frequency is high, so that averaged over all loci the synthetic will probably have higher variance. If the parent lines and synthetic are selected in closed populations of the same size for a long period of time the selection limit is expected to be higher for the synthetic than for the mean of the two pure lines. This relationship holds for both additive and completely dominant genes at all frequencies but the effects of linkage and epistasis are being ignored. However we are making the basic assumption that the traits under selection are influenced by a large number of loci, so there are only small differences in mean gene frequencies between the alternative populations. If there are wide differences in mean initial frequency the synthetic could have higher initial variance than the best line, yet never catch up with it under continued selection. But this would seem unlikely, especially as one population may have genes segregating which are absent from another. In general however, we lack concrete evidence and have an unsound basis for making practical decisions.

In the Institute of Animal Genetics in Edinburgh a relevant experiment with *Drosophila melanogaster* has been started by LOPEZ-FANJUL. Response to selection for sternopleural bristle number is being measured in two populations (Kaduna and Pacific) from different locations which have been maintained in cages in the laboratory for many years, and in synthetics formed from crosses between them. The initial performance of the two populations is almost exactly the same, but Pacific shows rather higher genetic variance and has responded somewhat more rapidly to selection. The cross shows no significant heterosis. With selection started from the F1 generation the synthetic has advanced at a rate intermediate between that of the parent lines. After allowing six generations of random mating without selection after the cross the heritability was estimated in another sample of the synthetic. Although a higher heritability value was obtained from the offspring-parent regression at this time, the subsequent selec-

tion response was no faster than in the parent lines. This result is rather hard to interpret, for one would expect an increase in genetic variance in F2 and later generations if there was negative linkage disequilibrium between the populations making the cross, but this should be accompanied by greater subsequent response. These results are as yet preliminary and the experiments are small. Nevertheless it is clear that the synthetic has little or no more additive genetic variance than the parent lines, which suggests that essentially the same loci are segregating in the two populations. More definitive conclusions will be possible when selection limits are reached. Unlike our domestic species these populations have no history of selection, so we should be cautious about making inferences from the *Drosophila* work.

E. — *Maintain alternative sire or dam lines.*

In addition to selecting in our chosen breeds A_1 and B_1, selection could be continued alongside in other populations, although their merit may be less at present. Of course the synthetic could be one of these. If rather different criteria were chosen for selection in these populations the programme would be much more flexible in that alternative breeds could be substituted as market demand and economic conditions change. The main disadvantage of this kind of scheme is that these potential substitute breeds have to be selected at almost the same rate as the ones already used, or they will gradually lag behind for the major traits and can never be utilised. Thus the breeding programme becomes much larger and more expensive. The same requirement has to be met for any breed which may be crossed into A_1 or B_1 in future years because it has some particularly valuable feature. Unless these breeds have performance near that of A_1 or B_1 the new synthetic A or B will be inferior. However there could be benefits from forming new synthetics if reproductive performance in A_1 or B_1 had deteriorated with inbreeding.

If our objective is to maximise gain over a long period of time, yet our facilities for maintaining animals under selection are limited, we have two distinct options. A synthetic can be formed immediately and selected as a single population. Alternatively the separate populations can be maintained as smaller populations, and each selected for a period before crossing and reselecting as a single larger population. ROBERTSON (1960) and MARUYAMA (1970) have shown that the same limit is obtained in either case. However the average rate of response will be higher if the synthetic is made initially since the subpopulations will become inbred more rapidly. But in the short term, in generations at least, our best strategy is probably to select in the highest ranking available breed or population.

ECONOMIC ASPECTS

Attempts have been made recently to evaluate breeding programmes in monetary terms using, in effect, the discounted cash flow procedure commonly employed in management accounting. The principles of the technique were first used in a genetic context by POUTOUS and VISSAC (1962) and subsequently by SOLLER, BAR-ANAN and PASTERNAK (1966). I shall give these in outline, and discuss their implications on alternative breed and cross bred improvement programmes.

Returns and costs incurred in any year are discounted back to some base, perhaps the year at which a decision is made to build a new testing station, or perhaps merely to the year at which a selection decision is made. For example, with an interest rate of 8 %, £ 100 invested now would realise £ 108 next year, £ $100 \times (1.08)^2$ the following year and so on. Thus £ 108 earned next year is equivalent to having only £ 100 now, or £ 1 obtained next year is worth £ $1/1.08$ = £ 0.926 now, and £ 1 earned 5, 10 or 20 years later is equivalent to £ 0.68, £ 0.46 and £ 0.21 earned now. With such an approach we can compute the aggregate benefits of selection response which are both permanent (at least in terms of changes in the traits) and cumulative. We can calculate either an overall „ profit " or the investment yield, which is the interest rate at which the scheme would just break even. Widely different programmes can be compared, or the returns from minor changes in selection procedure, involving relatively small extra expenditure, can be evaluated. Of course many simplifying assumptions need to be made, and it is difficult or impossible to take account of unforeseen changes in economic conditions. Such risks can be hedged to some extent by adopting discount rates considerably in excess of current interest rates. For example an estimated yield of 20 % evaluated over a period of only 15 years might be considered necessary before undertaking a programme. Especially when high discount rates are used the returns made in early years are weighted very heavily; it is this property of the procedure which has most relevance to our discussion of crossbreeding, for with large animals any programmes undertaken are likely to be of a long-term nature.

Consider the merit of maintaining synthetics or other substitute breeds of lower initial performance, but with the hope that they will eventually surpass the present superior population. No returns are obtained from this synthetic until the nucleus herd has reached the level of that of the superior breed, itself under selection, and until the population has been multiplied and progeny marketed. We considered earlier an example with beef cattle where the synthetic would require 6 years to catch up. We have to add to this, say, 2 years for bulls to mature and have progeny by A.I. and another 2 years before progeny are slaughtered, making a total of 10 years in all. At 10 years the discount factor is 0.46 if the rate is 8 %, and 0.16 if it is 20 %. Further, the extra returns after this period come only from the *increased* gain of the synthetic over the original breed, although only one selected population, the synthetic, now has to be maintained.

Using the same arguments it becomes difficult to justify maintaining several pure breeds or strains as potential substitutes. These must be selected at rates near those of the current commercial populations if they are ever likely tobe competitive, whether or not the objectives in the schemes are exactly the same. The costs of maintaining and selecting these populations will inevitably be considerable. Our rather simplified arguments lead us, therefore, to the conclusion that almost all our attention should be devoted to improving the breeds or crosses which are currently best. However a breeding organisation or country committing itself to such a scheme is vulnerable to a change in consumer demand or an exhaustion of genetic variance. But no scheme runs entirely in isolation, for there are competitors or other countries running similar programmes. These offer the best potential source of new variation!

LIMITATIONS

In conclusion a few comments should be made about the limitations of the analysis. In the first place it has been idealistic, and has by-passed many practical difficulties and limitations imposed by existing breeding systems, and by breeders' and farmers' prejudices. For example there may be resistance to use of what is clearly the best breed, or there may be legislation, as in Britain, to prevent the use of crossbred bulls. Even within the theoretical framework many simplifying assumptions have been made. In particular, interactions have been ignored both at the genetic level, between loci, and at the applied level, between environments. Nor has any general solution been given, but this is not possible with our current state of knowledge. There is clearly considerable need for greater understanding of the genetics of the major quantitative traits in our domestic species.

Reçu pour publication en octobre 1970.

RÉSUMÉ

ASPECTS THÉORIQUES DU CROISEMENT

Une discussion des méthodes d'utilisation des races de bovins et ses croisements, tenant compte des aspects génétiques et économiques, est présentée. L'essentiel de la théorie des comparaisons entre les races et leurs croisements est analysé mais, surtout, on a développé l'amélioration des croisements actuels.

Sans doute, une « population synthétique » aura plus de variabilité génétique et les limites de la sélection seront portées plus loin que celles des races qui la composent. Cependant, il s'écoulera souvent plusieurs années avant que cette « population synthétique » ne surpasse la meilleure race sous sélection continue. Pour cette raison les résultats économiques d'un tel procédé restent douteux.

En dépit de la marge de manœuvre que l'on a, en conservant plusieurs races, il faut les sélectionner continuellement si on veut qu'elles restent compétitives. Ainsi on peut attendre de meilleurs résultats par une sélection plus intensive des meilleures races existantes. Théoriquement, des schémas de sélection, basés sur les croisements, comme la sélection récurrente réciproque, ne permettent guère de faire progresser la sélection des gros animaux où les qualités de croissance et de carcasse sont importantes.

REFERENCES

DICKERSON G. E., 1969. Experimental approaches in utilising breed resources. *Anim. Breed. Abstr.*, **37**, 191-202.

FEWSON F., JAKUBEC V., 1971. Gewinnfunktionen-ein Hilfsmittel für die Planung von Gebrauchskreuzungen beim Schwein. *Ann. Génét. Sél. Anim.*, **3**, 000.

HILL W. G., 1970. Theory of limits to selection with line-crossing. In Kojima, K., *Mathematical topics in population genetics*, 210-245. Springer, Heidelburg.

JACKSON N., JAMES J. W. Comparison of three Australian Merino strains for wool and body weight. II. Estimates of between stud parameters. *Aust. J. Agric. Res.* (in press).

JAMES J. W., 1966. Selection from one or several populations. *Aust. J. Agric. Res.*, **17**, 583-589.

LERNER M., 1954. *Genetic homeostasis*. Oliver and Boyd, Edinburgh.

MARUYAMA T., 1970. On the fixation probability of mutant genes in a subdivided population. *Genet. Res.*, **15**, 221-225.

MOAV R., 1966. Specialised sire and dam lines. *Anim. Prod.*, **8**, 193-202, 203-211, 365-374.

MOAV R., HILL W. G., 1966. Specialised sire and dam lines. IV. Selections within lines. *Anim. Prod.*, **8**, 375-390.

POUTOUS M., VISSAC B., 1962. Recherche théorique des conditions de rentabilité maximum de l'épreuve de descendance des taureaux d'insémination artificielle. *Ann. Zootech.*, **11**, 233-256.

ROBERTSON A., 1960. A theory of limits in artificial selection. *Proc. Roy. Soc. Lond.*, B, **153**, 234-249.

SMITH C., 1964. The use of specialised sire and dam lines in selection for meat production. *Anim. Prod.*, **6**, 337-344.

SOLLER M., BAR-ANAN R., PASTERNAK M., 1966. Selection of dairy cattle for growth and milk production. *Anim. Prod.*, **8**, 109-120.

19

Size of experiments for breed or strain comparisons

by

William G. Hill

Proc.Working Symposium Breed Evaluation and Crossing Experiments, Zeist, 1974.

SIZE OF EXPERIMENTS FOR BREED OR STRAIN COMPARISONS.

William G. Hill, Institute of Animal Genetics, Edinburgh, EH9 3JN, Scotland.

Summary

The optimum size of experiments for comparing breeds are discussed from the viewpoint of maximising the net monetary returns over the cost of the test, as it is argued that possible monetary returns should not be ignored in test design. Two examples are given : where a new breed is compared to a standard breed in current use, which is replaced if the new breed performs better in the test; and where a large number of breeds or strains are compared in a random sample test. Formulae are given for calculating the optimum number of replicates. Although these depend on estimates of several monetary and physical parameters, they are moderately robust against quite large changes in assumptions.

Introduction

In recent years there have been many investigations into the financial returns from breeding programmes and how the magnitude and time scale of the returns and costs affect the choice of design. These studies have primarily been concerned with programmes for the improvement of breeds or strains by selection within them. Until the best available population has been identified there is obviously scope for improvement by selecting between populations, and this has the potential advantage that it can be achieved rapidly. With a few exceptions, for example that of the Meat and Livestock Commission (Scientific Study Group, 1971; D.E. Steane, personal communication) there has been much less study of returns from breed substitution. The particular problems which merit analysis are: the potential superiority of new breeds in monetary terms; the time scale of substitution, which affects the returns when discounted forward at realistic interest rates; the fraction of the industry likely to change breed (market penetration); what further improvements will probably accrue from subsequent selection within the new breed; and last, but chronologically the first criterion, the design, size and costs of the breed comparison programme. It is the last of these aspects which will be discussed in this paper.

The problems of experimental design are primarily discussed in statistical texts. However, the subject has been introduced here in a monetary context to emphasise that, especially in the large scale experi-

ments often required for farm animals, financial considerations predominate. Experiments designed to compare breeds of animals are often very expensive: the costs including those for new housing, perhaps with facilities for individual feeding, arrangements for breeding together with the purchase of stock for testing and recurrent costs for recording, individual feeding and carcass dissection. Thus any experiment should be as small as possible, but compatible with being sufficiently accurate that correct decisions are usually made as a result of the test, because large returns from an industry can accrue from breed substitution. Although data can often be collected cheaply from field records, these are unlikely to be satisfactory for traits, such as feed conversion efficiency, which are difficult to measure, or on new breeds which are not widely used.

The question we should ask of a proposed experiment to compare two or more breeds is : what is the expected monetary return, not what is the probability that the null hypothesis of no difference between them is rejected, as in classical theory? Inevitably, the difficulty with such an approach is that we have to determine, or guess, the values of even more variables than if financial aspects are excluded. But although any specific application is likely to be imprecise, this does not imply that no analysis should be undertaken.

Several important aspects of breed comparisons will not be discussed in this paper: for example, the problem of sampling individuals from the breed such that the sample mean is unbiassed, or specifically is an unbiassed estimate of the mean of the population which might then be used subsequently, is ignored; the optimum family size to use in the tests, bearing in mind that it is usually more expensive to sample 100 progeny from 20 sires than from 10 is not considered, but an analysis has been made by Connolly (1974); and the problems of combination of information on several traits are not pursued, the argument being based on the idealised assumption that a single trait, or a known index, is a sufficient description of merit. Only two examples are discussed in any detail: that where one new breed is available to be tested and which, if superior, would then replace the current breed; and where there are a large number of equivalent strains or breeds available from which one is to be chosen, such as in a random sample poultry test. Although the methodology does not apply only to breed comparisons, examples will be given in that context.

Comparison between a new and current breed

Let us assume that a particular breed, say A, is currently being used and an alternative, B, is available on which we have no prior knowledge. We wish to compare these breeds (either as pure or cross breds, the argument is not substantially affected) and make the appropriate choice between them on the basis of the test. In a standard (Neyman-Pearson) approach to the size of the experiment we would specify two probabilities: of accepting a null hypothesis of no difference between the breeds when one exists (Type II error), and of rejecting this null hypothesis when true (Type I error). The necessary size of the experiment is then proportional to the square of the minimum difference in true performance between the breeds we wish to be able to detect. A more simple approach is to consider only the type I error. This was done by Comstock and Winters (1942) in an early discussion of the design of breed comparisons, but can be criticised on statistical grounds in that the method gives only a 50% chance of detecting a difference when one exists. But is there any reality to a null hypothesis that two breeds have the same mean, even though they may have the same potential role in the industry? Surely the breeds are different, the question is which one should we choose? Bechhofer (e.g. Bechhofer et al., 1968), for example, has long argued that the null hypothesis has no relevance in such ranking problems.

In this kind of experiment a decision theory approach seems more useful. In the simplest model a difference, d, is observed between breeds A and B, and if B turns out to be sufficiently superior to A that the costs and problems of introduction (importation and multiplication of breeding stock) can be covered, A is replaced by B, otherwise A is retained. Whilst this may be a somewhat unrealistic model it should serve as illustration. In practice there may be some prior information on the new breed, perhaps on some traits (e.g. growth rate in cattle) but not others (e.g. feed conversion efficiency), or in a different environment. The presentation appears rather formal, but this is only to show how one particular approach can be used. In practice the methodology, or at least its general principles, can be adapted to other, perhaps more realistic situations.

The single stage procedure described here is almost certainly not optimal, and more elaborate two stage procedures have been discussed by Grundy et al. (1956). They consider whether the new process (i.e.

breed) should be accepted or rejected after the first test or whether further experimentation should be undertaken, and if so, how large the extra trial should be. This and other sequential procedures have the attraction that they are more efficient in use of facilities having a high capital cost, but produce results later, perhaps at an indeterminate time. With breed experiments where it may take a year or more to complete an individual test, sequential tests may take too long, or at least produce later and thus more heavily discounted returns. In practice, however, expensive test facilities are unlikely to be used only once: new breeds will be tested or further tests carried out on the same breeds (but it may be difficult to stop the industry making wrong decisions based on partial results).

Returning now to the single stage problem, let us assume that: the returns to the industry from each unit of improvement are W (some problems of computing W are discussed later); the cost of replacing the previously used breed by the new one would be V, which covers importation, multiplication of stock etc.; the overhead costs of the test, which do not depend on its size are H, the additional cost for each pair of test places is C, and there are n pairs of test places (i.e. n each of the new and existing breed); the standard deviation of a single observation on the difference between A and B in performance on the last is σ; and the true, but unknown, difference in performance, B-A, between them is δ. If the observed difference in the test is d, the value of a breed substitution predicted from that test is Wd, and if Wd exceeds V, the cost of the changeover, B is introduced; but if Wd< V the old breed, A, is retained.

The return, R, (or negative loss in the decision theory context) expected from any single test thus depend on the probability, P (Wd> V), that the new breed is accepted,

$$R = (W\delta - V)\, P(Wd > V) - nC - H. \qquad (1)$$

The probability of acceptance can be calculated under the usual normal distribution assumptions, and depends on the sample size. An important requirement of any test to be established is that, for likely values of δ, (which may be based on prior information) the expected return, R, is positive.

Equation (1) can be generalised by defining $\varepsilon = (\delta - V/W)/\sigma$ as the standardised superiority of the new breed above that necessary to cover replacement costs, and $\gamma = C/W\sigma$ as the cost of a replicate as a proportion of the expected returns from an improvement equal to one standard deviation of

one replicate. Eq. (1) reduces to

$$R = W\sigma \{\varepsilon F(\varepsilon \sqrt{n}) - \gamma n\} - H \qquad (2)$$

where $F(\cdot)$ denotes the standardised normal distribution function. The effect on returns of changes in the size of the test thus depends only on ε and γ, and the returns are maximised by the value of n which satisfies

$$f(\varepsilon \sqrt{n})/(\varepsilon \sqrt{n}) = 2\gamma/\varepsilon^3 \qquad (3)$$

where $f(\cdot)$ denotes the standardised normal density function. Solutions to (3) are given in Figure 1 and depend on the absolute value of ε, not its sign.

In Figure 1 we see that the optimum test size always increases as γ, the cost of testing relative to the value of improvement, decreases. However when γ is very small the optimum number of replicates is higher when the difference in performance between the breeds is small than when it is large, and vice versa when γ is large. The explanation is that with large ε, (standardised difference between the breeds less replacement costs) few replicates are necessary to establish with high probability that ε is positive, so the correct decision is made; whereas with small ε the large amount of testing necessary is not justified in terms of returns when the cost of testing is high.

Let us put the methods in context by considering an example, perhaps not a very realistic one, but other values can be substituted. Assume that two breeds of beef cattle are to be compared for feed conversion efficiency (f.c.e.). The capital costs (using approximate figures based on those kindly supplied by D.E. Steane of the Meat and Livestock Commission) are about £440 per animal housed (excluding grants) and the running costs about £40 per animal, net over returns. To save additional discounting computations on the fixed costs, let us attach an annual depreciation and interest charge to the capital of £60 per animal, making the total costs £100 per animal, or C = £200 per pair. Assuming the only trait of interest is f.c.e. to 400kg live weight, each extra unit change in f.c.e. of feed costing £50/ton is worth £20 per animal. Imagine the results were to be used for a population of 100000 animals in which, if the breed substitution were recommended, it would take, say, 10 years to complete, starting 2 years after testing and occurring linearly. Using a 20 year period of evaluation and a 10% discount rate the discounted returns would be the same as on 466000 animals now. Hence $W = £9.3 \times 10^6$. Taking the standard deviation of

f.c.e. as 0.6 (e.g. mean of 6 and coefficient of variation of 10%), that for the difference between a pair of animals would be $0.6\sqrt{2} \doteq 0.85$. Assuming the breed substitution cost, involving purchase of 10 bulls and 50 cows is V = £60000 and the test overheads (manager, computing etc.) are H = £10000, eq. (1) becomes

$$R = (9.3 \times 10^{6}\delta - 60000)P - 200n - 10000,$$

and in (2) $\varepsilon = 1.18\delta - 0.007$, $\gamma = 2.53 \times 10^{-5}$. Thus if the new breed had a superiority in the range of $\delta = 0.05$ to 0.4 (or roughly inferior by that amount), covering most of the values of interest, the optimum of $\sqrt{n}$ is in the range 7 to 17, i.e. 50 to 290 pairs should be tested (from Fig. 1). Whilst this is quite a wide range, in a conventional Neyman-Pearson formulation and allowing type I and type II errors each of 5% the optimum is $\sqrt{n} = 2.79/\delta$, giving values of n ranging from 50 to 3120 pairs for δ in the range 0.05 to 0.4. The other important question is how robust are optima against changes of assumptions, particularly those involving W, which is the most speculative? For W increased or reduced by a factor of 10, and with $\delta = 0.2$, the optimum for n rises to 220 pairs or falls to 60 pairs from its present value of 140. Thus rather large changes in assumptions can be made without corresponding changes in the optimal design being necessary.

It is also informative to plot expected returns against sample size for different values of the true breed difference, δ. This is done in Fig. 2 for the example described above. Since both the fixed cost, H, and the costs of a single replicate are small, relative to W, the returns are positive for all positive values of δ of interest; in other words, the test is likely to be justified. Notice that the curves for returns against n are very flat topped at small values of δ, and also for higher n values with δ large. Thus considerable changes can be made in the design with little loss in efficiency. In other situations where the values of γ (testing cost vs value of improvement) are much smaller, the curves drop off more rapidly at high values of n since too large testing costs are incurred.

Comparison among several alternative breeds

As a contrast to the case where two breeds are being compared let us consider the situation where there are many alternative breeds or strains which are competitors for the same market. An example of this would be in a random sample test for egg laying poultry where stocks (which may, of course, be crosses) are entered by many breeders. To simplify the problem

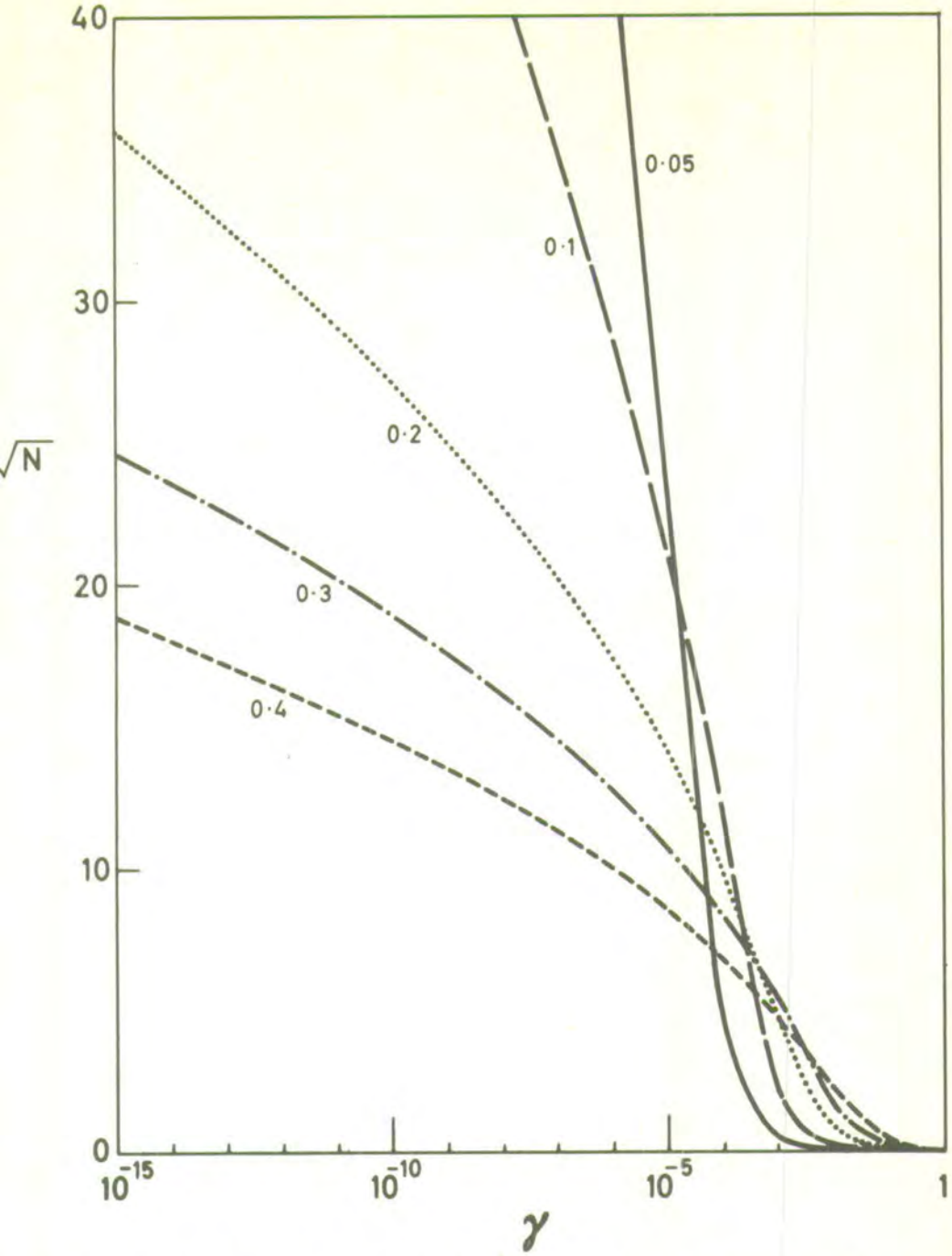

Fig. 1. Optimum number of pairs (n) to test in a two breed experiment as a function of standardised test costs (γ) and standardised difference between populations (ε).

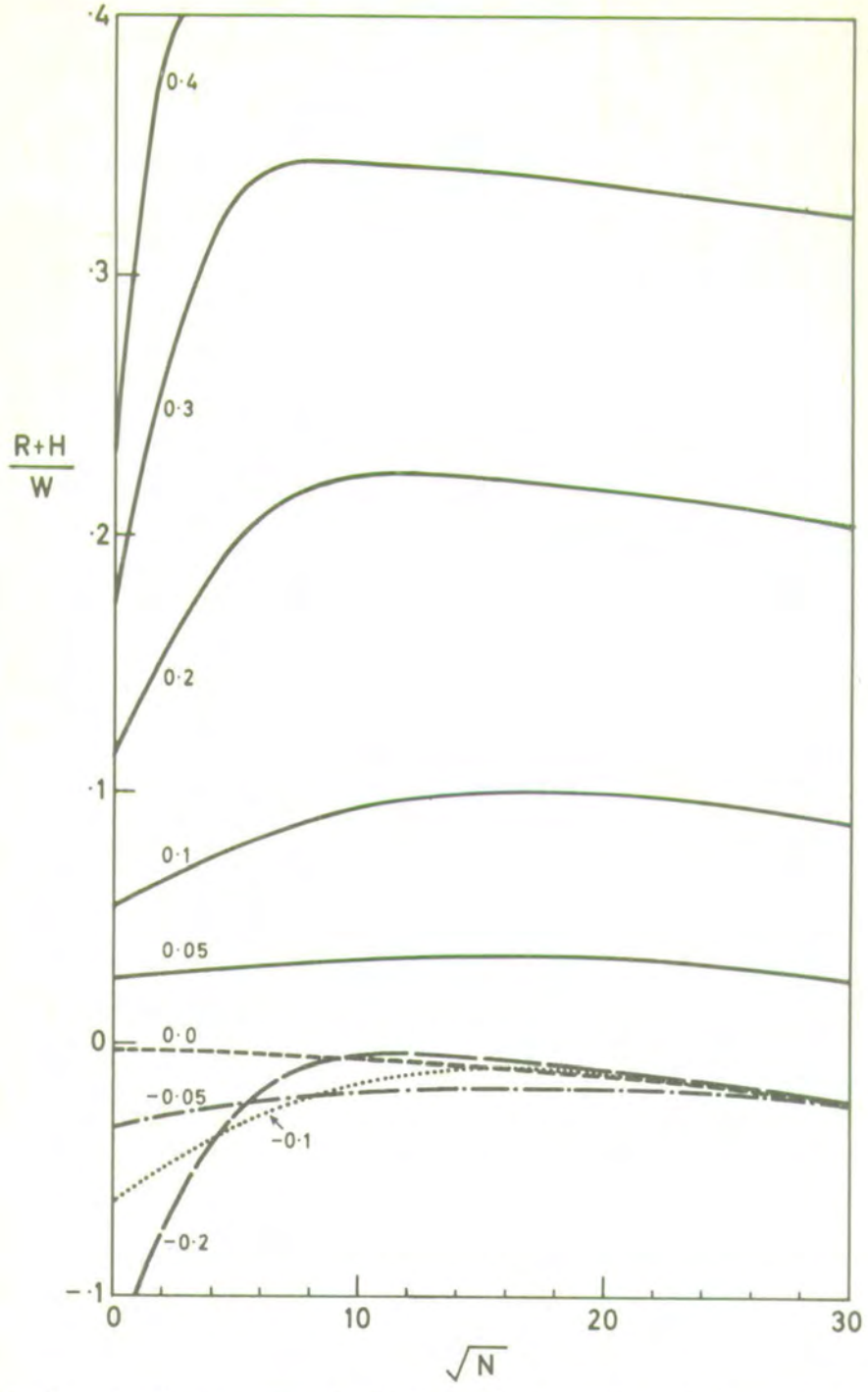

Fig. 2. Expected returns, as (R+H)/W, in the beef cattle example as a function of the difference between the breeds (δ) and number tested (n).

we assume there is no recognised standard breed against which the others are being compared, and that our objective is just to pick the best of those under test.

The number of animals to test in such programmes can be tackled by standard methods using probabilities of type I and type II errors. Recently, Connolly (1974) has made calculations of numbers required on this basis, when different costs are incurred for individual testing of progeny and use of different numbers of their sires. There are several other methods. Becker (1961) discussed the size of a trial in terms of the probability that the best strain ranks first, following the work of Bechhofer, reviewed by Bechhofer et al. (1968). Taylor (1974) has used a minimax approach to find the optimum allocation of test spaces and number of breeds to select, such that the chosen group exceeds the true mean. In none of these cases, however, are expected returns from the scheme taken into account; for example, it does not matter greatly if the second best breed is picked if it is only marginally poorer than the best. An alternative approach to this problem (first suggested to me by Alan Robertson some years ago) is to compute expected genetic gains from the breed testing programme by putting it in the context of a selection programme, in which the expected gain equals the selection differential among breeds multiplied by the regression of future on test performance. This approach lends itself to an analysis of monetary returns, which can be computed from the expected gain and counterbalanced against increased costs incurred by improving the accuracy of the test with increased replication.

The same notation is used as before, except that σ^2 and n now refer to the number of replicates of each stock (not of pairs) and k strains are tested. The tested strains are assumed to be taken from a population of strains with variance σ^2_b between strains (they may be a sample or the whole population). The observed variance of test means is thus $\sigma^2_b + \sigma^2/n$ and the regression of true on test performance is $\sigma^2_b/(\sigma^2_b + \sigma^2/n)$. The observed superiority of the best strain and its impact on the industry can be summarised by the selection differential in standard deviations, i. (The gain as a result of the test depends to a large extent on the selection differential applied by the industry. As anyone involved with a breeding scheme knows, it is not always the highest scoring animals that are selected; sometimes no selection is practised at all.) The returns from the test are therefore

$$R = Wi\sigma^2_b(\sigma^2_b + \sigma^2/n)^{-\frac{1}{2}} - nkC - F, \qquad (4)$$

where the response is measured relative to the mean of the strains under test. If this mean differs from that of populations currently being used commercially, a correction to (4) is necessary. However the difficult problem of evaluating experiments with several strains which have jointly to be compared with a existing standard strains will not be discussed. The test is only justified when the expected returns, from (4) or after some modification, are positive using reasonable values of parameters.

If other parameters in (4) are known, differentiation with respect to n gives the optimal size of the test. The solution is given by

$$n^2(\sigma^2_b + \sigma^2/n)^{3/2} = \sigma^2_b\sigma^2 \, Wi/2kC \qquad (5)$$

which has to be obtained numerically. This gives a simple upper limit to n of

$$n < (Wi\sigma^2/2kC\sigma_b)^{\frac{1}{2}}. \qquad (6)$$

A similar generalisation of these formulae can be achieved as in the previous section by some reparametrisation. The essential parameters are the ratio of variances, σ^2_b/σ^2 (which corresponds to ε), the cost of measurement relative to possible returns, $nC/W\sigma^2$ (which corresponds to γ), and the ratio of selection differential to the number of strains on test, i/k, which can not exceed 0.28 for normally distributed populations. This generalisation will not be pursued further.

In the example used before, W = £9.3 x 10^6, σ = 0.6, C = £100. Assuming k = 8 breeds were tested, then with normally distributed breed means, the highest ranking breed would have an expected selection differential of 1.4. To allow for some lack of use of the text results, let us take i = 1. With small differences among the breeds, say σ_b = 0.1. Then (6) gives n < 145 and (5) gives n = 120. Substituting n = 120 in (4) shows the expected return to be £7.1 x 10^5 approximately, a large figure since test costs are small.

Although again faced with the necessity to make very many assumptions, these calculations of numbers are moderately robust, although less than in the two breed case considered previously. It is seen in (6) that of the important parameters which are likely to be difficult to estimate, n is a function of the square root of W, i and σ_b. The distribution of the true means of stocks enters the assumptions, but primarily though its effect on i.

In a plant breeding context there are likely to be many new varieties produced which can be tested, and thus extra returns from increasing the selection intensity and accuracy of selection. There have therefore been many studies on optimal sequential screening methods (Finney, 1958). With animals, except perhaps inbred lines of poultry, the number of strains available for testing is likely to be small, so the changes in selection intensity, i, that can be effected are likely to be due primarily to increasing the use of the results by the industry. It would be interesting to see market research studies of the effect of random sample tests on the use of egg laying stocks, such that the selection differential achieved in practice could be determined.

Discussion

The analysis may be thought unduly theoretical, for it is clear that in any practical situation it will be difficult to obtain satisfactory estimates of parameters, in particular, W, the returns for unit improvement. Nevertheless, the object of this paper has been to focus attention on the costs/returns relationship involved in designing breed comparisons, rather than to give recipes on designs which are too straightforward and might be misapplied. It is clear that no breeder would undertake a test without considering the cost involved, in the words of Finney (in discussion of Grundy et al., 1956) these are questions of "internal economy", but the potential returns, or questions of "external economy" often seem to be thought out in less detail. Of course, with most assumptions of returns and possible breed differences, it will turn out that large profits are to be expected from breed tests. The problem may then be one of optimal allocation of resources to alternative breeding (or other) programmes. Then it may be necessary to predict the marginal discount rate of the test so that money can be spent in an optimal way. Certainly some calculations on costs and possible returns seem desirable, even without adequate knowledge of the necessary parameters.

Some of the over-simplifications of the analyses described here need emphasis. Throughout it was assumed that the standard deviation, σ, was known without error, and not subject to modification. As mentioned previously, it can be changed by altering the family size. There does not seem to be a simple balance between the extra costs which can be incurred to compensate for a reduction in variance due to increasing the number of sires, but it might merit study. Similarly, we have not included any

genotype-environment interaction component. If there is prior information to suggest there is an interaction, a test would have to be replicated in several locations. Decisions could either be made within environments in which case each single test would have to stand on its own; or the mean performance over environments could be used as a criterion. In the latter case the interaction would appear as an additional source of variance and the reduction in variance over the hypothetical average environment balanced against the costs of testing in several locations.

The problems of computing the returns per unit improvement, W, have been largely ignored. The example given illustrated many of the difficulties; assumptions have to be made about the size of the market, the degree and rate of penetration of the new breed if successful, how long it is used and the marginal value of the trait several years from now. Also some of these variables, such as the rate of penetration, may be affected by the differences demonstrated in the test. Finally, we have discussed a single analysis; in practice there is likely to be some prior information and tests giving equivocal results may be repeated.

Acknowledgement

I am grateful to Marjorie McEwan for computational assistance.

References

BECHHOFER, R.E., KIEFER, J. and SOBEL, M. 1968. Sequential identification and ranking procedures. Univ. Chicago Press.

BECKER, W.A. 1961. Comparing entries in random sample tests. Poultry Sci. 40, 1507-1514.

COMSTOCK, R.E. and WINTERS, L.M. 1942. Design of experimental comparisons between lines of breeding in livestock. J. Agr. Res. 64, 523-532.

CONNOLLY, J. 1974. Economic and statistical optimisation of beef breed comparisons. Proc. Genetic Operations Research Workshop, Trinity College, Dublin (Abstr.)

FINNEY, D.J. 1958. Plant selection for yield improvement. Euphytica 7, 83-106.

GRUNDY, P.M., HEALY, M.J.R. and REES, D.H. 1956. Economic choice of the amount of experimentation. J. Roy. Stat. Soc. B18, 32-55.

MEAT AND LIVESTOCK COMMISSION. 1971. Beef improvement. Scientific Study Group Report.

TAYLOR, St.C.S. 1974. Multibreed theory. Proc. Genetic Operations Research Workshop, Trinity College, Dublin. (Abstr.)

The possible use of superovulation and embryo transfer in cattle to increase response to selection

by

Roger B. Land and William G. Hill

Anim. Prod. 1975, 21: 1–12

THE POSSIBLE USE OF SUPEROVULATION AND EMBRYO TRANSFER IN CATTLE TO INCREASE RESPONSE TO SELECTION

R. B. LAND
ARC Animal Breeding Research Organisation, Edinburgh EH9 3JQ
AND
W. G. HILL
Institute of Animal Genetics, Edinburgh EH9 3JN

SUMMARY

The possible use of superovulation and embryo transfer in selection programmes in cattle is investigated theoretically, in terms of both rates of response and inbreeding.

In a selection programme for growth rate, it should be possible to achieve about twice the response of a conventional performance testing programme, so that 400-day weight, for example, could be increased by 16 rather than 9 kg per year.

The improvement of reproductive performance by the use of laparoscopy to measure the natural ovulation rate of animals over several oestrous cycles followed by superovulation of selected animals is investigated. The rate of progress is dependent upon the incidence of twin ovulations in the base population and is unlikely to exceed 0·6% per year unless the initial frequency is 8% or more.

INTRODUCTION

THE cost of beef production is particularly dependent upon the growth characteristics of the animals for slaughter and the number of animals reared per dam. Genetic selection for either of these components of productivity is restricted by the low reproductive rate of cattle, which necessitates a long generation interval and, in females, a low selection intensity. The development of embryo recovery and transfer techniques, however, indicates that it may be possible to increase the reproductive rate of females following superovulation (Rowson, 1971), and to obtain several calves from a single adult female (cow) in one ovulation or over a period of a few months. In addition to the commercial use of this technique to facilitate the rapid multiplication of superior (or at least novel) imported breeds, it could also be used to increase female selection intensity within closed populations. We shall consider its application to the improvement both of characteristics of the growing animal and of reproductive performance.

The traits of growth rate and feed conversion efficiency of young cattle have a high heritability, should respond readily to selection on individual performance, and schemes based on a small closed herd have been suggested (Meat and Livestock Commission (MLC), 1971).

The improvement of the second component, reproductive performance, would increase the number of animals available for rearing, and although there are management difficulties which may be associated with increasing the twinning rate of cattle (e.g. calf mortality, increased calving interval and decreased milk production) the extra calf production in beef cattle for suckling

may outweigh the potential disadvantages. Selection to increase the frequency of twinning is restricted by the low natural incidence and the sex-limited expression of twinning. The incidence of twinning is around 0·5 to 1% at first calving and 2 to 4% in older cows of the Friesian breed. It has a low heritability and repeatability, possibly of the order of 4% and 6% respectively (Bowman and Hendy, 1970; Hendy and Bowman, 1970; Donald, 1974; Bar-Anan and Bowman, 1974), although the recent analyses of Johansson, Lindhé and Pirchner (1974) suggest even these may be optimistic estimates. Some herds of the large French breeds, however, are reported to have higher twinning rates, 5 to 7% (F. Menissier, personal communication), although the highest incidence quoted by Ortavant and Thibault (1970) is 4·6% for the Simmental. The simplest scheme based on conventional reproduction in a genetically closed herd entails selecting all replacement males (bulls) from cows which give twins at one or more calvings, with little selection among female replacements. With a heritability of twinning of 4% and a low repeatability, around 6%, the rate of response in twin births could be in the region of 0·1 to 0·15%/year, which seems trivial. This could be increased by progeny testing bulls and selecting them on their daughters' second and later calving records (the twinning rate being too low at first calving) but would require for effective operation a large field testing programme, a significant proportion of that required for a national milk improvement programme. A high twinning herd probably could be established by collecting superior females from the national herd, and initially using bulls which had been progeny tested nationally for milk production and shown to have daughters with a high incidence of twinning. Such cows or bulls could not be used subsequently, for unless the selected herd is genetically isolated it would not improve faster than the national herd, so progress would be limited to the first generation. We shall therefore consider the identification of superior individuals by detecting twin ovulations in a series of ovarian examinations by laparoscopy (Mariana, 1969), or possibly rectal palpation, in successive oestrous cycles, rather than simply observing twin births, followed by the superovulation of such individuals. The assumption implicit in the method is that ovulation rate is the main limitation of twinning, as discussed by Ortavant and Thibault (1970).

In this paper we suggest possible designs and indicate the rates of progress which might be achieved using the physiological aids of superovulation and laparoscopy in schemes to improve either traits of the growing animal or the incidence of twinning. We can only give a general impression of what might be achieved, for we do not have good estimates of all the necessary parameters. We shall, however, use conservative estimates of these parameters, and future improvements in techniques in reproductive physiology may make further increases in selection response possible. Any scheme involving superovulation and transplantation needs access to facilities which have a high initial capital expenditure in a surgery and equipment, and incurs a large annual charge for skilled labour; but the paper will be restricted to the genetic problems and the number of animals required.

DEFINITIONS AND ASSUMPTIONS

There is little published evidence on which to base an estimate of the likely yield from superovulation. Rowson, Moor and Lawson (1969)

transferred 109 fertilized eggs recovered from 42 donor cows superovulated with pregnant mare's serum (PMS), an average of 2·6 embryos per donor, but they do not say if this was the total number collected. Of those transferred in their best treatment, 12 of 13 cows became pregnant but at least 12 of the cows were each given two eggs. In a further study, 13 of 18 cows given one embryo in each uterine horn became pregnant, and 22 of 36 eggs were represented as calves or as implantations (Rowson, Lawson and Moor, 1971). Foote and Onuma (1970) reported that the yield of 4·4 cleaved ova per cow by Scanlon, Sreenan and Gordon (1968) was one of the more successful experiments, but no reference was made to embryo viability. More recently an average of 3·3 pregnant recipients following a single superovulation treatment has been reported by a Canadian group (R. B. Church, personal communication). If embryo survival is independent of the number transferred it would be reasonable to expect from the above data that 70% of transferred embryos would survive, to give two to four calves per collection per donor. Assuming that repetition of this procedure is possible, even though at present it is not normal practice, a cow may therefore 'produce' six to eight calves over a period of 1 to 2 months.

The following terms are introduced to facilitate the comparison of conventional and potential selection programmes:

p_{δ}, p_{φ}: proportions of males and females selected,
i_{δ}, i_{φ}, $\bar{\imath}$: corresponding standardized selection differentials, and their mean (assuming a large population),
L_{δ}, L_{φ}, $\bar{L}$: mean age of parents when progeny are born (generation interval),
x: mating ratio (female mates per male),
s: proportion of calves surviving from early embryo and not culled for traits other than those under primary selection, *viz.* growth rate or ovulation rate (taken as 0·8),
k: the number of concepta per donor,
C: the total number of cows in the herd, donors plus recipients,
ΔF: the rate of inbreeding/year.

If the schemes using superovulation are to produce responses consistently higher than conventional schemes they must utilize closed populations. The rate of response per year equals $(\bar{\imath}/\bar{L})\,h^2\sigma$ for a trait with heritability h^2 and phenotypic standard deviation σ (Falconer, 1960). The heritability and standard deviation are characteristics of the selected trait, but $\bar{\imath}/\bar{L}$ depends on the structure of the population and can be used as a prediction of the relative merits of alternative population selection and replacement schemes. Thus we shall compare different structures in terms of $\bar{\imath}/\bar{L}$, and refer to this as the '*annual selection intensity*'. Alternative schemes also incur different rates of inbreeding, a high rate being likely to involve long term disadvantages of depressed performance, particularly in reproductive traits, and loss of genetic variation. The rate of inbreeding ΔF for a herd of given structure is inversely proportional to its size. The costs of maintaining a herd using embryo transplantation are roughly proportional to the total number of donor and recipient cows, C. Thus we have combined inbreeding and cost considerations, measured in terms of the total number of cows, in the parameter $C\Delta F$, and refer to it as the '*herd rate of inbreeding*'. For example, a value of $C\Delta F$, of 0·26 implies an annual rate of inbreeding of 0·26% in a

herd of 100 cows. If N_{δ} and N_{φ} are the number of males and females, respectively, entering the donor herd each year (or the whole herd in the conventional scheme) the rate of inbreeding is taken as

$$\Delta F = (1/N_{\delta}+1/N_{\varphi})/8\bar{L}^2$$

(Hill, 1972). (The generation interval, $\bar{L}$, appears as a squared term in this equation since the number of breeding animals is proportional to $\bar{L}$ and the number that enter per year, and the annual rate of inbreeding, is $1/\bar{L}$ of that per generation.) There are assumed to be no differences between families in viability and fertility or in family size as a result of artificial selection, so the formula is likely to underestimate the real rate of inbreeding.

The incidence of twin ovulations has been reported from rectal palpation studies to be 13·1 and 5·4% in American Holsteins by Kidder, Barrett and Casida (1952) and Labhsetwar, Tyler and Casida (1963), respectively. Despite the difference between these estimates and the difficulty of identifying twin ovulations accurately by rectal palpation, they do indicate that the incidence is higher than that of twin births. Preliminary laparascopy data indicate incidences of 12, 3 and 7% for twin ovulations in Friesian, Hereford and Simmental, respectively (R. B. Church, personal communication). Indeed, if one assumes that the conception rate of cattle is a measure of the probability of successful fertilization and implantation, and that this is independent of the number of eggs shed, then the square of this probability indicates the proportion of twin ovulations likely to be represented as twin births. Using the non-return rate as a crude upper estimate of conception rate (63% from Wijeratne and Stewart, 1971) the incidence of twin ovulations may be two to three times that of twin births. In the absence of estimates of the heritability of twin ovulations we shall assume it to be equal to that of twinning. This may give an underestimate of value, for the estimated heritability of ovulation rate in mice exceeds that of litter size (31 *v.* 15%) (Falconer, 1963; Land and Falconer, 1969). Finally, we assume that increases in ovulation rate will lead to increases in the incidence of twinning, an assumption supported by the increase in twinning of cows calving following the transfer of two eggs in cattle, referred to above, and by analogy from the critical role of the ovary and the ovulation rate in the determination of the litter size of the sheep, discussed by Land (1974).

IMPROVEMENT OF CHARACTERS OF THE GROWING ANIMAL

Schemes using superovulation will be compared to schemes such as that described by the MLC (1971) which rely on the conventional reproductive performance of females. Both schemes are based on the selection of males and females on a performance test at 1 to 1¼ years of age.

For the *conventional scheme* we have assumed that cows and bulls have their first progeny when 2 years old, that 90% of cows calving in any year survive to the following year, and that young and mature cows have the same calving rate. (Small differences in these parameters make little difference to the conclusions.) In Table 1 computed values of the annual selection intensity ($\bar{\imath}/\bar{L}$) are shown for a range of mating ratios and ages at which males and females are culled. The annual selection intensity is rather insensitive to the choice of structure, and with customary mating ratios

(1 : 20 to 1 : 40) is about 0·4. The rate of herd inbreeding, however, is sensitive to the choice of mating ratio (Table 1).

For the *superovulation scheme* the main factor which influences the rate of progress is the number of calves reared from each donor cow. We assume that cows in the donor herd can be superovulated when they are around 15 months of age, and the eggs transferred to recipient cows, with each receiving one egg. If the embryo does not survive, a second transfer is made to the recipient and we assume all recipients are fertile. Bulls in the donor herd

TABLE 1

Predicted annual selection intensity ($\bar{i}/\bar{L}$) *and herd rate of inbreeding* ($C\Delta F$) *in a conventional beef cattle performance testing scheme with* C *cows. Results are given for different replacement policies of bulls and cows, shown as the ages of animals when their progeny are born*

Mating ratio (x)	10		20		40		80		10		20		40		80	
Ages of bulls (yr)	2	2–3	2	2–3	2	2–3	2	2–3	2	2–3	2	2–3	2	2–3	2	2–3
	$\bar{i}/\bar{L}$								$C\Delta F$							
Ages of cows (yr)																
2–4	0·29	0·33	0·37	0·39	0·43	0·45	0·49	0·49	0·26	0·39	0·47	0·72	0·88	1·40	1·70	2·76
2–5	0·32	0·36	0·39	0·41	0·45	0·46	0·50	0·51	0·23	0·34	0·41	0·63	0·75	1·21	1·45	2·37
2–6	0·33	0·36	0·39	0·41	0·45	0·46	0·50	0·50	0·21	0·30	0·36	0·56	0·66	1·06	1·25	2·07
2–7	0·33	0·36	0·39	0·40	0·44	0·45	0·48	0·49	0·19	0·28	0·32	0·50	0·58	0·94	1·10	1·84

are also assumed to have all their progeny when about 2 years of age, so that the generation interval is 2 years, but a lower mating ratio is used than in the conventional herd to reduce inbreeding, for the scheme is not so dependent on selection in males. A range of values were considered for the number of eggs (k) successfully transferred per donor. In practice wide variability between individual egg collections is likely, but variation between donors can be reduced by adjusting the number of collections and discarding excess eggs from large individual collections; thus an average figure of eight eggs should be achieved. Allowing for subsequent culling, with 80% surviving, the number of animals of each sex available for selection ($\frac{1}{2}ks$) should be approximately three. The proportions of females and males selected for growth rate are then 1/3 and $1/3x$, respectively ($p_{♀} = 2/ks$, $p_{♂} = 2/ksx$). Predicted annual selection intensities and rates of herd inbreeding are given in Table 2. The composition of the donor herd can be described in terms of the total cow population (C) and the number of recipients

TABLE 2

Predicted annual selection intensity ($\bar{i}/\bar{L}$) *and herd rate of inbreeding* ($C\Delta F$) *in a beef cattle performance testing scheme using superovulation, in terms of the number of progeny per donor available for selection* (ks) *with* s = 0·8

Mating ratio (x)	1	2	4	8	16	1	2	4	8	16
	$\bar{i}/\bar{L}$					$C\Delta F$				
ks										
2	0·00	0·20	0·32	0·41	0·49	0·22	0·33	0·55	0·98	1·86
3	0·27	0·41	0·51	0·60	0·67	0·30	0·45	0·74	1·34	2·52
4	0·40	0·52	0·61	0·69	0·76	0·38	0·56	0·94	1·69	3·19
6	0·55	0·65	0·73	0·81	0·87	0·53	0·80	1·33	2·39	4·52
8	0·64	0·73	0·81	0·88	0·94	0·69	1·03	1·72	3·09	5·84
10	0·70	0·79	0·87	0·93	1·00	0·84	1·27	2·11	3·80	6·75

pregnant per donor (k), assuming no spare recipients. This gives $N_{\varphi} = C/(1+k)$ females and $N_{\sigma} = C/x(1+k)$ males entering the donor herd each year, so that when $\bar{L} = 2$, $C\Delta F = (x+1)(k+1)/32$. Furthermore, the scheme assumes random mating in the donor herd. If, however, the scheme were run with a restricted annual calving season the 2-year generation interval could lead to the genetic division of the donor herd into two groups each giving progeny in alternate years. Some animals would therefore have to be retained and used for a second year to maintain the genetic unity of the population. Alternatively two independent selection programmes could be integrated so that each used the same recipient herd in alternate years, when annual rates of progress (proportional to $\bar{\imath}/\bar{L}$) would not be affected, but the rate of inbreeding would be doubled.

Assuming six offspring can be obtained per donor, with a mating ratio of eight, almost double the response is predicted for the superovulation relative to the conventional scheme. This results from the reduced generation interval and, dependent on the female age structure in the conventional herd, the increased selection intensity. Taking as typical figures a heritability of 0·5 and a standard deviation of 40 kg of live weight as 13 months for a beef breed of large body size (MLC, 1971), an annual selection intensity of 0·45 in a conventional scheme corresponds to an annual response of 9 kg, whereas $\bar{\imath}/\bar{L} = 0{\cdot}8$ for the superovulation scheme corresponds to 16 kg/year.

This basic design could be modified in several ways; for example a reduction in the number of recipient cows, and thus in the number of operations, could be achieved by transferring two eggs to each recipient, but there would be a loss in selection intensity among females since most of those born co-twin to a male would be sterile. When one egg is transferred to each recipient with a mating ratio of 4, $\bar{\imath}/\bar{L} = 0{\cdot}73$ and $C\Delta F = 1{\cdot}33$ (Table 2). If, instead, two eggs were transferred to each recipient with a survival rate per embryo of 0·75 and if all females born co-twin to a male discarded at birth, the equivalent figures would be $\bar{\imath}/\bar{L} = 0{\cdot}65$ and $C\Delta F = 0{\cdot}90$. By raising the mating ratio to 8 with two eggs per recipient the predictions become $\bar{\imath}/\bar{L} = 0{\cdot}73$ and $C\Delta F = 1{\cdot}60$. On these calculations the transfer of two eggs per recipient appears to offer no advantages, and subsequent analyses are limited to the transfer of single eggs.

The calculations of rates of progress have been based on selection on individual performance, but in the conventional scheme half-sib family information and in the superovulation scheme full-sib family information could be included in an index. Since full sibs are all reared by different dams, in the superovulation scheme there should be no common environmental effects among full sibs. With a heritability of 0·5 and a half-sib family size of 32 ($x = 40$, $s = 0{\cdot}8$) in the conventional scheme an index would be 6·5% more efficient than individual selection. In the superovulation scheme with a full-sib family size of six the increase in efficiency would be 9%. Index selection, however, would also lead to an increase in the rate of inbreeding.

IMPROVEMENT OF REPRODUCTIVE PERFORMANCE

In the scheme to be evaluated the natural ovulation rate is observed by laparoscopy, or possibly by rectal palpation, over successive oestrous cycles, enabling several observations to be made in a short period of time. As in the case of growth we shall first describe and consider a basic programme

and then discuss some of the possible alternatives. The animals or families showing most multiple ovulations are selected, super-ovulated and embryos transferred to recipient cows as in the growth rate scheme described previously. In view of the four-fold increase in twinning rate (see Introduction) observed between first and second parities, which might just be an age effect and possibly mediated by a decrease in embryonic mortality, we suggest, however, that females are initially mated naturally at around 1·2 years to calve at 2 years. Then for a period from about 2·2 to at most 3 years of age, up to 10 ovulations are observed by laparoscopy before animals are selected, superovulated and their embryos transferred, so that they will then be an average of 3·5 years old when their progeny are born. Some animals with a high twin ovulation rate could be selected on the basis of fewer observations, but we base our calculations on an equal number on each animal, and have chosen 5 or 10 as examples; with 10 a slight increase in generation interval could be required, but this has not been considered in the calculations.

The additional parameters required are as follows:

q: proportion of twin ovulations,
n: mean number of females tested in a full sib family ($n = ks/2$, approximately),
m: number of laparoscope observations per cow, so mq is the average number of twin ovulations per cow over m observations,
r, h^2: repeatability, heritability of twin ovulations at a single oestrus,
A: breeding value of selected animals, so the annual rate of response to selection is equal to $(\bar{\imath}\bar{A})/\bar{L}$, where $(\bar{\imath}\bar{A})$ is averaged over the two sexes.

If the heritability and repeatability of twin ovulations at a single oestrus are assumed to be similar to those of twin births, then taking Bowman and Hendy's (1970) figures of 4% for heritability and 6% for repeatability, the heritability of the mean of $m = 5$ or 10 ovulations rises to 16% or 26% respectively (Table 3), to give a trait of intermediate heritability. Also, although the incidence (q) of twinning in a single oestrus may be low, say 4%, the incidence of cows with at least one twin ovulation over 10 ovulations approaches 40%, implying that about one-third of cows would show at least one twin ovulation and individual selection could be practised.

The computational problem of predicting response is that the trait has an all-or-none expression, so selection differentials cannot be computed accurately from the normal distribution. More general results can be obtained using the normal approximation so we have compared the predicted selection differentials, and thus response, from it with that obtained directly from the all-or-none case, using the most extreme situation where almost all animals have either none or only one twin ovulation ($mq \ll 1$). For female family sizes (n) of three the ratios of predicted selection differentials, all-or-none/normal are 0·41, 0·57 and 0·81 for average frequencies of twin ovulations observed per cow (mq) of 5%, 10% and 20% respectively; with a heritability of 4% and a repeatability of 6% and at higher incidences the assumption of no more than one twin ovulation per cow breaks down and the normal approximation is preferred. Therefore, if $mq > 20\%$, approximately, predictions based on the normal distribution are satisfactory and will be used. In practice sufficient ovulations might be observed so that animals with twin ovulations are selected (i.e. when mq just exceeds $1/n$, i.e. $2/ks$).

For a range of possible parameter values the heritabilities and standard deviations of the mean number of twin ovulations on individual animals are given in Table 3. These are also combined in the Table to give the expected breeding value (A) of a group of individual animals whose average performance exceeds that of the population mean by one standard deviation (i.e. $i = 1$), together with breeding values of full sib males in families where three females are recorded, and whose family performance differs by one standard deviation. Thus if a selection differential of one standard deviation were applied to select females on their individual performance and males on their full sibs' performance, the expected breeding value of selected females and males would be 2·00% and 1·54% respectively, for the typical parameter

TABLE 3

Heritability (h^2_m) *and standard deviation* (σ_m) *of the mean ovulation rate of* m *laparoscope observations for different values of heritability* (h^2), *repeatability* (r) *and incidence* (q) *of twinning on single observations. The expected breeding value* (A_1) *of individual animals which exceed the mean performance of the population by one phenotypic standard deviation* (*i.e.* i = 1) *and the expected breeding value* (A_s) *of male full sibs in families where three females are measured, and whose mean performance exceed the population mean by an average of one phenotypic standard deviation are also given*

					σ_m%†				A_1%				A_s%			
h^2%	r%	m	h^2_m%†	q%:	2	4	8	16	2	4	8	16	2	4	8	16
2·5	5	5	10		6·9	9·6	13·3	18·0	0·71	1·00	1·38	1·87	0·59	0·82	1·14	1·54
—	—	10	17		5·3	7·5	10·3	14·0	0·92	1·29	1·78	2·41	0·74	1·03	1·42	1·93
4	6	5	16		7·0	9·8	13·5	18·3	1·12	1·57	2·18	2·94	0·90	1·26	1·75	2·37
—	—	10	26		5·5	7·7	10·6	26·0	1·43	2·00	2·77	3·74	1·10	1·54	2·13	2·88
5	10	5	18		7·4	10·4	14·4	19·4	1·32	1·85	2·56	3·46	1·06	1·48	2·04	2·76
—	—	10	26		6·1	8·5	11·8	26·3	1·61	2·25	3·11	4·21	1·24	1·73	2·40	3·24

† $\sigma^2_m = q(1-q)[1+(m-1)r/m]$, $h^2_m = mh^2/[1+(m-1)r]$.

values of $h^2 = 4\%$, $r = 6\%$, $m = 10$ and $q = 4\%$. If the females were selected on an index of individual and family performance the expected breeding value of selected females would be increased from 2·00 to 2·23%.

The figures in Table 3 show the important effect which the current incidence of twinning has on possible progress. Furthermore, if a model of a threshold trait is used to compute heritabilities at different incidence levels, higher values of heritability are likely to be associated with higher incidences (Robertson and Lerner, 1949). For example, an increase in incidence from 4% to 8% would increase heritability from 4% to 6%.

There are several alternative selection schemes which could be run. One comprises individual selection of females and random replacement of young males, giving parental ages of 3·5 for females, 2 for males and a generation interval of 2·75 years. With no selection applied to males the response is independent of the mating ratio and a function of the mean family size. The rate of response is listed here in terms of full sib family size for comparison with Table 2:

Family size ($2n = ks = 2/p_{♀}$):	3	4	6	8	10	12
Annual selection intensity ($\bar{\imath}/\bar{L}$):	0·099	0·145	0·198	0·231	0·254	0·273

The responses can be computed as a product of these values and those given in Table 3, and some predicted responses are listed in Table 4. Since no

selection is practised on males, the rate of inbreeding can be reduced by using a mating ratio as low as one with no loss of response and little cost (since the males could be slaughtered after use at marketable age) and by choosing one replacement male from each family. Typical values for rates of inbreeding are given in Table 4. These correspond to the values in Table 2, but are reduced to take account of the increased generation interval and the replacement of one male from each half-sib family. With the parameters used in Table 4 the expected progress is 0·31%/year for an incidence of 4% and 5 observations; with 10 observations this rises to 0·40% and to 0·55% if the incidence is 8%. For an annual rate of inbreeding of 0·5% a total of only about 40 cows would be required if the mating ratio was unity, of which about 20 would require the laparoscopies each year and 7 would be selected and superovulated. Taking one bull to four cows ($x = 4$), 2·5 times as many cows would be required.

TABLE 4

Predicted annual change in the incidence of twin ovulation (%) *and herd rate of inbreeding* (CΔF) *when* ks = 2n = 6 *for* (*a*) *selection on individual female performance alone, and* (*b*) *when males are also selected on the mean performance of their three sibs,* ($h^2 = 4\%$, r = 6%)

		(*a*) Females only			(*b*) Males and females	
L (years)		2·75			3·5	
Mating ratio (*x*)		1	4	16	4	16
CΔ*F*		0·21	0·53	1·79	0·43	1·48
m	*q*%					
10	2		0·28		0·42	0·53
5	4		0·31		0·47	0·60
10	4		0·40		0·59	0·75
10	8		0·55		0·82	1·03
10	16		0·74		1·10	1·39

An alternative scheme would be to select males on their full sib family mean, giving a generation interval of 3·5 years. To reduce inbreeding and costs only one young male per family would be retained for possible mating, so the proportion selected on their sisters' performance would be $1/x$. The rate of inbreeding is roughly 0·3 of that given in Table 2 to allow for the increased generation interval. Examples are given in Table 4 for $n = 3$ sisters recorded, and in a typical case ($x = 4$, $m = 5$, $q = 4\%$) the annual improvement is predicted to be 0·47% compared with 0·31% for males chosen randomly, a marked change. Because the generation interval is increased the expected annual rate of inbreeding is lower with sire selection for the same mating ratio, but when sires are not selected pair mating can be used.

These alternatives could be further improved in two ways. First, the work associated with the identification of superior individuals could be reduced if laparoscopy was only continued until a twin ovulation was recorded, up to a specified maximum number. In this way, with a low repeatability for the incidence of twinning the mean number of observations taken on each selected cow would be halved. Secondly, it might be possible to programme the scheme so that calves born to the first, natural, mating could be used for breeding, thereby increasing the annual selection intensity.

Combination of schemes for growth and reproductive rate

In beef cattle breeds for individual or multiple suckling husbandry systems it may be desirable to improve both growth rate and reproductive performance (although improved growth rate may carry a penalty in increased mature size). If it were desirable, the two programmes suggested could be combined. Males would be selected on their growth rate and females on their ovulation rate, so that $L_{\delta} = 2, L_{\varphi} = 3{\cdot}5$ and $\bar{L} = 2{\cdot}75$ years. Assuming that the traits are uncorrelated, the rate of response for reproductive performance would be that given previously (Table 4a). The annual selection intensities for growth rate with three female offsprings per donor and a mating ratio of 4 and 16 would be 0·33 and 0·44, respectively. These give values of predicted response up to one-half of that when selection is practised solely for growth rate on both sexes (Table 2), and are similar to those which can be achieved in a selection scheme for growth rate using conventional breeding techniques (Table 1).

DISCUSSION

The incorporation of superovulation and embryo transfer into a programme of selection in cattle for traits of the growing animal should enable the rate of response to be approximately doubled. To achieve this improvement without increasing the rate of inbreeding would require a total of two or three times as many cows as in a closed scheme using conventional reproduction. If the life of a selection experiment or programme was put at 20 years and the maximum tolerable inbreeding 10%, or 0·5%/year, a total herd of about 500 cows would be needed, assuming 8 mates/sire, but a reduction of the scheme to 300 cows with 4 mates per sire would be expected to increase the rate of response by 90% relative to the conventional scheme. Transferring two eggs to each recipient would reduce the number of recipient cows and operations by approximately 50%, but the number of operations on donors would be unchanged, and the rate of response would be reduced by approximately 10%. The choice of scheme would therefore depend on the relative costs both of egg recovery and insertion and of maintenance of donor and recipient cows.

To assess a superovulation scheme for growth traits, we list the total annual requirements, assuming a total (C) of 500 cows with one egg per recipient and approximate figures for a conception rate of 0·75, eight concepta per donor and six calves per donor suitable for selection on growth rate; this implies about 7 bulls, 55 donor and 445 recipient cows. These are:

Operations on donors: up to 3 operations on 55 cows, say 150;
Operations on recipients: on average 1·33 operations on 445 cows/year (to allow for non-viable ova), say 600;
Testing accommodation for growth rate: up to 400 animals (leaving 330 after culling);
Young cow and bull accommodation (mating ratio of 8): 70 animals;
Replacements in recipient herd (20% year): 90 cows.

These figures do not make any allowance for the selection of synchronous recipients.

It is clear that such a programme would be expensive, but if bulls bred in the scheme were used widely, say in artificial insemination, they could

generate a large return. For example, from Hill (1971), an extra 1 kg in the live weight of 200 000 crossbred calves at slaughter could be worth a total of about £180 000 to the British beef industry when discounted at 15% over 20 years. Clearly substantial costs could be carried by a scheme increasing response by 7 kg/year.

The usefulness of selection for twinning is harder to assess for the successful commercial exploitation of repeated twinning has not been demonstrated, although it could well prove to be profitable in a suckler system. The rate of response to selection for ovulation is difficult to estimate and dependent on the incidence in the base population. Furthermore, the proportion of twin ovulations that lead to twin births is not known. If, however, we accept the use of data from twin births, rates of response in the region of 0·3%/year, or only 6% in 20 years represent a small absolute improvement. The acceptability of such a scheme may therefore depend on the availability of a base population where the incidence of twinning or at least twin ovulations was 8% or more, giving an increase of 0·5 to 0·7%/year. Exploitation of semen from bulls whose daughters are known to have a high incidence of twinning, of breeds with a high spontaneous incidence and of severe selection of the initial cows might enable such a population to be established. The incidence might then be 10% or even 15%, and the subsequent progress possible is indicated in Table 3 using an incidence of 16%. Either then, or if current heritability estimates prove too low or if the incidence of twin ovulation is found to be double that of twinning, a scheme using laparoscopy and superovulation of selected females within a closed population might become feasible. Also, although small increases in the incidence of twinning may not in themselves be of practical value, such increases may render the population more amenable to the induction of twinning by treatment with exogenous hormones (P. Mauléon, personal communication).

The other situations where superovulation might prove a useful tool in improvement programmes are with dairy cattle and sheep. It is less obvious how the technique could be incorporated into the dairy bull progeny testing situation and make a very marked improvement to progress, and there could be greater costs associated with loss of production following operations on heavily lactating cows. In sheep the reproductive rate and hence selection potential on females is already much higher in many breeds than in cattle and therefore a superovulation scheme is less attractive than with beef cattle; laparoscopy may still be desirable however (Hanrahan, 1974).

We have tried to illustrate the feasibility of accelerated selection for growth rate and for reproductive performance, rather than to try to propose definite schemes, or to confuse the illustration by the presentation of too many detailed alternatives. We do not doubt that the present proposals could be further refined, for example by including progeny testing. We have, however, used conservative estimates of the variables on which our estimates are based, especially in the case of selection for twinning. Despite the absence of accurate basic information, we feel that we have demonstrated the advantages of physiological aids to artificial selection in domestic animals and indicated how further advances in reproductive technology may increase the rate of genetic progress. Specifically, although non-surgical transfer of eggs would reduced the work involve, increases in embryo yield by improvements in superovulation techniques or development of non-surgical recovery would enable the utilization of superior females to be increased.

ACKNOWLEDGEMENTS

We thank Monsieur F. Menissier, Dr I. Wilmut and Dr R. B. Church for their useful suggestions, and many colleagues for their comments.

REFERENCES

BAR-ANAN, R. and BOWMAN, J. C. 1974. Twinning in Israeli-Friesian dairy herds. *Anim. Prod.* **18**: 109–115.

BOWMAN, J. C. and HENDY, C. R. C. 1970. The incidence, repeatability and effect on dam performance of twinning in British Friesian cattle. *Anim. Prod.* **12**: 55–62.

DONALD, H. P. 1974. Twinning in cattle. *A.B.R.O. Annual Report*, pp. 27–32. H.M. Stationery Office, London.

FALCONER, D. S. 1960. *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh.

FALCONER, D. S. 1963. Quantitatively different responses to selection in opposite directions. In *Statistical Genetics and Plant Breeding* (ed. W. D. Hanson and H. E. Robinson). *National Academy Sciences—National Research Council, Washington. Publ. No.* 982, pp. 487–490.

FOOTE, R. H. and ONUMA, H 1970. Superovulation, ovum collection, culture and transfer. A review. *J. Dairy Sci.* **53**: 1681–1692.

HANRAHAN, J. P. 1974. Ovulation rate as a selection criterion for fecundity in sheep. *Proc. 1st Wld Congr. Genet. appl. Anim. Prod.*, Vol. 3, pp. 1033–1038. Madrid.

HENDY, C. R. C. and BOWMAN, J. C. 1970. Twinning in cattle. *Anim. Breed. Abstr.* **38**: 22–37.

HILL, W. G. 1971. Investment appraisal for national breeding programmes. *Anim. Prod.* **13**: 37–50.

HILL, W. G. 1972. Estimation of genetic change. I. General theory and design of control populations. *Anim. Breed. Abstr.* **40**: 1–15.

JOHANSSON, I., LINDHÉ, B. and PIRCHNER, F. 1974. Causes of variation in the frequency of monozygous and dizygous twinning in various breeds of cattle. *Hereditas* **78**: 201–234.

KIDDER, H. E., BARRETT, G. R. and CASIDA, L. E. 1952. A study of ovulations in six families of Holstein-Friesians. *J. Dairy Sci.* **35**: 436–444.

LABHSETWAR, A. P., TYLER, W. J. and CASIDA, L. E. 1963. Analysis of variation in some factors affecting multiple ovulation in Holstein cattle. *J. Dairy Sci.* **46**: 840–845.

LAND, R. B. 1974. Physiological studies and genetic selection for sheep fertility. *Anim. Breed. Abstr.* **42**: 155–158.

LAND, R. B. and FALCONER, D. S. 1969. Genetic studies of ovulation rate in the mouse. *Genet. Res. Camb.* **13**: 25-46.

MARIANA, J. C. 1969. A technique for the *in-vivo* examination of ovaries in the cow. *Annls Biol. anim. Biochim. Biophys.* **9**: 657–659.

MEAT and LIVESTOCK COMMISSION. 1971. *Beef Improvement*. Scientific Study Group Report. Meat and Livestock Commission, Bletchley, Bucks.

ORTAVANT, R. and THIBAULT, C. 1970. Reasons for obtaining twin births in cattle and procedures followed. *Annls Biol. anim. Biochim. Biophys*, **10**: Supplement 1, 1–19.

ROBERTSON, A. and LERNER, I. M. 1949. The heritability of all-or-none traits: viability of poultry. *Genetics* **34**: 395–411.

ROWSON, L. E. A. 1971. The role of reproductive research in animal production. *J. Reprod. Fert.* **26**: 113–126.

ROWSON, L. E. A., LAWSON, R. A. S. and MOOR, R. M. 1971. Production of twins in cattle by egg transfer. *J. Reprod. Fert.* **25**: 261–268.

ROWSON, L. E. A., MOOR, R. M. and LAWSON, R. A. S. 1969. Fertility following egg transfer in the cow: effect of method, medium and synchronisation of oestrus. *J. Reprod. Fert.* **18**: 517–523.

SCANLON, P., SREENAN, J. and GORDON, I. 1968. Hormonal induction of superovulation in cattle. *J. agric. Sci., Camb.* **70**: 179–185.

WIJERATNE, W. V. S. and STEWART, D. L. 1971. Population study of abortion in cattle with special reference to genetic factors. *Anim. Prod.* **13**: 229–235.

(*Received* 7 *August* 1974)

21

Effect of sampling errors on efficiency of selection indices

I. Use of information from relatives for single trait improvement

by

Jill Sales and William G. Hill

Anim. Prod. 1976, 22: 1–17

EFFECT OF SAMPLING ERRORS ON EFFICIENCY OF SELECTION INDICES

1. USE OF INFORMATION FROM RELATIVES FOR SINGLE TRAIT IMPROVEMENT

JILL SALES AND W. G. HILL
Institute of Animal Genetics, Edinburgh EH9 3JN

SUMMARY

An analysis is undertaken of the effect of errors in estimates of parameters, particularly the intra-class correlations, on the response from selection for one trait using an index of individual together with full- and/or half-sib family records. A distinction is drawn between the response (R) possible with use of the optimum index, that predicted ($\hat{R}$) and that achieved (R^*) with an index which uses the sample estimates of the parameter values.

It is found that the loss of efficiency (R^*-R) using sample estimates is very small even for estimates far from the correct value. $\hat{R}$ is more sensitive to errors, particularly of the heritability and phenotypic variance estimates. Since the latter also appear in the prediction of response from individual selection, errors in predicting the relative responses from index and individual selection are small.

Expected values of the proportional loss in response,

$$L = (E(R^*)-R)/R,$$

are approximately proportional to the variance of the estimate of intra-class correlation. It is shown that in practice initial experiments with 20 or so families may be sufficient to get average proportional losses down to less than 1%.

INTRODUCTION

INFORMATION on the performance of relatives can be incorporated into a selection index with the individual's own performance and used to increase genetic improvement. This information may be on one or more traits. The selection index is a linear weighted combination of observed measurements, constructed so as to maximize genetic gain. Although originally proposed in an animal context for combining information on several traits (Hazel, 1943) the same principles apply for combining information from several individuals (Henderson, 1963). For single traits, the combination of individual and full- or half-sib records was discussed by Lush (1947) and the incorporation of both full- and half-sib records by Osborne (1957).

In order to construct an index, estimates of genetic parameters are required. These may be obtained from a sample of data from the same population, or, if that is lacking, perhaps from data in the literature on similar populations in similar environments. The optimum response will only be obtained if the index is constructed using the precise parameter values; if the estimates are in error some efficiency will be sacrificed. Several

studies have been undertaken of the resultant loss in efficiency (Harris, 1963, 1964; Pig Industry Development Authority, 1965; Mao, 1971). These have primarily been concerned with cases where genetic merit and observations are based on several traits.

This study was undertaken to look in detail at indices based on measurements on single traits on an individual and (usually) his collateral relatives. The problem is taken in two parts: first, what is the effect on response when specific, incorrect, estimates of parameters are used; and secondly, how is the expected response affected by the variance of the estimates of parameters used to construct the index and thus by the number of animals from which data are taken to compute these estimates? Following Harris (1964), comparisons are made between progress (R) using the optimum index, progress ($\hat{R}$) predicted from the index using parameter estimates and the actual progress (R^*) which will be achieved when the index computed from parameter estimates is used in the population (these are ΔH, $\hat{\Delta} H$ and $\Delta H'$, respectively, in Harris's notation). Some preliminary results have been published previously (Hill, 1974).

THEORY

Examples of specific applications are given subsequently, but for single traits the general theory for selection indices (Henderson, 1963) reduces to the following. Information is available from k sources on each individual (e.g. $k = 2$ for individual performance and full-sib family mean). An index

$$I = \sum_{i=1}^{k} b_i x_i = \boldsymbol{b}'\boldsymbol{x}$$

is constructed where $\boldsymbol{b}$ is a vector of k index weights and $\boldsymbol{x}$ a vector of k observations (symbols are defined in Table 1). The response in the breeding value (A) of the trait, expressed as a ratio of the selection differential in standard deviations, is

$$R = \rho_{AI}\sigma_A = \boldsymbol{b}'\boldsymbol{G}(\boldsymbol{b}'\boldsymbol{P}\boldsymbol{b})^{-\frac{1}{2}}, \tag{1}$$

where ρ_{AI} is the correlation between the breeding value of the trait in the individual and the index, $\sigma^2{}_A$ is the additive genetic variance, $\boldsymbol{P}$ is the $k \times k$ variance-covariance matrix of the observations $\boldsymbol{x}$, and $\boldsymbol{G}$ is the vector of k covariances of observations with breeding value of the individual. Response (1) is maximized when the index satisfies

$$\boldsymbol{b} = \boldsymbol{P}^{-1}\boldsymbol{G}, \tag{2}$$

giving

$$R = (\boldsymbol{G}'\boldsymbol{P}^{-1}\boldsymbol{G})^{\frac{1}{2}}. \tag{3}$$

The weights $\boldsymbol{b}$ computed from (2) can be multiplied by any constant without changing the ranking of individuals or response. However the standardized values given by (2) are convenient in that the regression of breeding value on the index is unity.

In any practical situation, only estimates $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{G}}$ of the parameters $\boldsymbol{P}$ and $\boldsymbol{G}$ will be available. The weights of the estimated index, $\hat{I}$, are usually taken as

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{G}} \tag{4}$$

TABLE 1

Definition of symbols (addition of ^ to any symbol denotes estimate, addition of ′ to any matrix denotes transpose)

A	breeding value
$\mathbf{x}$	vector of observations (number of observations = k, usually 2)
$\mathbf{b}$	vector of index weights
I	index value
ρ_{AI}	correlation of index and breeding value
$\boldsymbol{P}$	variance-covariance matrix of observations
$\boldsymbol{G}$	vector of covariances between observations and breeding value
h^2	heritability
t	intra-class correlation of sibs
σ^2	phenotypic variance (σ^2_A variance of breeding value)
B	mean square between families
W	mean square within families
r	coefficient of relationship of family members ($r = \frac{1}{4}$ for half-sibs, $\frac{1}{2}$ for full-sibs)
s	number of families
n	number of progeny in each family
E, V	statistical expectation (mean), variance
R	response using the optimum index (assuming a selection differential of one standard deviation)
$\hat{R}$	response predicted using estimates of parameter values
$R^\star$	response achieved using estimates of parameter values

$D = \frac{1}{2}\partial^2 R^\star/\partial \hat{t}^2$, i.e. $E(R^\star) = R + DV(\hat{t})$

$L = [E(R^\star) - R]/R$, proportional loss

R_1, $\hat{R}_1$, $R_1^\star$ optimum, predicted and achieved responses from single trait selection

Both full- and half-sib families present:

d	number of dams per sire, n number of progeny per dam
t_s	intra-class correlation of half-sibs, t_d correlation of full-sibs within half-sibs

and the predicted progress is then given by

$$\hat{R} = (\hat{\boldsymbol{G}}'\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{G}})^{\frac{1}{2}}. \tag{5}$$

The actual progress achieved using $\hat{I}$ is given by

$$\begin{aligned} R^\star &= \operatorname{cov}(A, \hat{I})[V(\hat{I})]^{-\frac{1}{2}} \\ &= \hat{\boldsymbol{b}}'\boldsymbol{G}(\hat{\boldsymbol{b}}'\boldsymbol{P}\hat{\boldsymbol{b}})^{-\frac{1}{2}} = \hat{\boldsymbol{G}}'\hat{\boldsymbol{P}}^{-1}\boldsymbol{G}(\hat{\boldsymbol{G}}'\hat{\boldsymbol{P}}^{-1}\boldsymbol{P}\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{G}})^{-\frac{1}{2}} \end{aligned} \tag{6}$$

from Harris (1963), and this must be less than or equal to R, the response obtained using the optimum weights.

Formulae (5) and (6) enable predicted and achieved progress to be compared with optimum progress for any specified set of parameter estimates. When the estimates are obtained from a sample of data on individuals from the same population, it is useful then to consider the expected values of $\hat{R}$ and $R^\star$, and their deviation from R over conceptual replicate samples of data. Since larger samples of data should give, on average, better estimates of parameters, the problem becomes one of specifying adequate sample sizes to obtain a reliable index.

METHODS

Estimates of $\boldsymbol{P}$ and $\boldsymbol{G}$ can be obtained from analyses of sib or offspring-parent data, or combinations of both. Precise values for expectations of

response have been obtained in this paper only for data from a balanced one-way classification of families and individuals within families. Then, for normally distributed observations, the between-family and within-family mean squares are distributed independently as chi-square multiplied by a constant factor. The expected values of the necessary parameters, and thus predicted and achieved responses were obtained by integrating numerically over the ranges of the mean squares using a modification of Simpson's rule for two variables.

Approximate results for a wider range of models have been obtained by using a Taylor's series approximation as outlined in the Appendix. For an index of, say, individual and full-sib family performance, the weights depend only on the intra-class correlation, t, of sibs (Lush, 1947). If an unbiased estimate $\hat{t}$ is available, the formula for $E(R^\star)$ for example, reduces to

$$E(R^\star) = R + \tfrac{1}{2}V(\hat{t})\left.\frac{\partial^2 R^\star}{\partial \hat{t}^2}\right|_{\hat{t}=t} \tag{7}$$

approximately, where the second derivative $\partial^2 R^\star/\partial \hat{t}^2$ is invariably negative when evaluated at t, since $R^\star \leqslant R$. Approximations such as (7) are based on the assumption that estimates of parameters do not depart too far from their true value and so are satisfactory only when sample sizes are fairly large. They do, however, have the advantage of generality and simplicity. For example, when the intra-class correlation is estimated from s half-sib families of specified size, $V(\hat{t})$ and thus the loss $E(R^\star)-R$ are inversely proportional to s (providing s is sufficiently large that terms in $s-1$, say, are well approximated by s).

Some results using Taylor's expansion have been checked by Monte Carlo simulation, assuming normally distributed observations.

RESULTS

(i) *Individual and full- or half-sib family mean performance*

Response for specific estimates of parameters. Probably the most common index application is where records are available on an individual and its full- or half-sibs. The variables used are

x_1 = deviation of the individual's observation from the family mean, and

x_2 = family mean.

(The alternative formulation, where x_1 is the individual's performance expressed as a deviation from the overall mean, gives the same responses). Then

$$\boldsymbol{P} = \frac{\sigma^2}{n}\begin{pmatrix}(n-1)(1-t) & 0\\ 0 & 1+(n-1)t\end{pmatrix}, \quad \boldsymbol{G} = \frac{h^2\sigma^2}{n}\begin{pmatrix}(n-1)(1-r)\\ 1+(n-1)r\end{pmatrix} \tag{8}$$

where σ^2 is the phenotypic variance, t is the intra-class correlation of sibs, h^2 is the heritability, n is the family size (including the individual) and assumed to be the same for all families, and r is the coefficient of relationship between sibs (0·25 for half-sibs, 0·5 for full-sibs) (see Table 1). If the parameters h^2, t and σ^2 are known without error, the index weights are, from (2)

$$b_1 = \frac{h^2(1-r)}{1-t}, \quad b_2 = \frac{h^2[1+(n-1)r]}{1+(n-1)t} \tag{9}$$

giving

$$b_2/b_1 = \{[1+(n-1)r](1-t)\}/\{[1+(n-1)t](1-r)\}, \tag{10}$$

and from (3)

$$R = \frac{h^2\sigma}{\sqrt{n}}\left\{\frac{(n-1)(1-r)^2}{1-t} + \frac{[1+(n-1)r]^2}{1+(n-1)t}\right\}^{\frac{1}{2}} = h^2\sigma\left\{1 + \frac{(n-1)(r-t)^2}{(1-t)[1+(n-1)t]}\right\}^{\frac{1}{2}} \tag{11}$$

compared with a response of $h^2\sigma$ from selection on individual performance alone, as shown by Lush (1947). If only estimates of the parameters are available $\hat{h}^2$, $\hat{t}$ and $\hat{\sigma}^2$ replace the corresponding parameters in (8)–(11), the substitution in (11) giving $\hat{R}$. Note that the relative weights in (10) depend only on $\hat{t}$ but not on $\hat{h}^2$. The actual progress is, from (6).

$$R^\star = \frac{h^2\sigma}{\sqrt{n}}\left\{\frac{(n-1)(1-r)^2}{1-\hat{t}} + \frac{[1+(n-1)r]^2}{1+(n-1)\hat{t}}\right\} \times \left\{\frac{(n-1)(1-r)^2(1-t)}{(1-\hat{t})^2} + \frac{[1+(n-1)r]^2[1+(n-1)t]}{[1+(n-1)\hat{t}]^2}\right\}^{-\frac{1}{2}}. \tag{12}$$

For (12) a positive value of $\hat{h}^2$ is used to compute the index; otherwise the phenotypically inferior animals are selected and $R^\star$ is negative.

Throughout the paper it is assumed that the same family structure is present in the initial sample as is used subsequently for calculating the index scores of individuals. So for an index based on the individual measurement and the mean of four half-sibs, say, the parameters are assumed to have been estimated from a half-sib analysis of variance with a family size of four.

The estimate of heritability ($\hat{h}^2$) is not critical except for predicting response. Assuming there are no environmental correlations among half-sibs, then in the case of half-sibs t may be estimated as the correlation within families and $\hat{h}^2$ taken as $4\hat{t}$. Estimating $\hat{h}^2$ from $\hat{t}$ is not so realistic in a full-sib structure where the correlation between full-sibs may be substantially affected by maternal environmental effects and dominance. In the following a distinction has been drawn between $2\hat{t}$ and $\hat{h}^2$ for full-sib families, whereas for half-sibs $4\hat{t}$ and $\hat{h}^2$ are regarded as equivalent.

In Figure 1, values of $R^\star$, the response achieved, are plotted for two family sizes and several different values of t against estimates $\hat{t}$ in a half-sib structure where the true heritability is assumed to be $4t$ and the phenotypic standard deviation equal to 1. It is seen that $R^\star$ is rather insensitive to the estimate of intra-class correlation, a range of 0·1 or more in $\hat{t}$ about the correct value t having little effect on the response. The predicted response, $\hat{R}$, is also shown for three possible values of $\hat{\sigma}$, and with the heritability estimate taken as $4\hat{t}$. Thus $\hat{R}$ is very sensitive to the value of $\hat{t}$, as would be the case with individual selection, since $\hat{R}$ is roughly proportional to $\hat{t}$ and also the estimate, $\hat{\sigma}$, of the phenotypic standard deviation. If h^2 were estimated elsewhere, $\hat{R}$ would be much less sensitive to $\hat{t}$ and although not shown in Figure 1, would not lead to negative values of $R^\star$ when $\hat{t}$ was negative. In Figure 2 values of $R^\star$ are given for full-sib families, with the estimate of heritability assumed to be positive for all values of t, and for purposes of illustration h^2 is taken equal to $2t$. The actual response $R^\star$ is again very

insensitive to $\hat{t}$, and here, since h^2 is known and assumed to be positive, R^* remains positive even when $\hat{t}$ is negative.

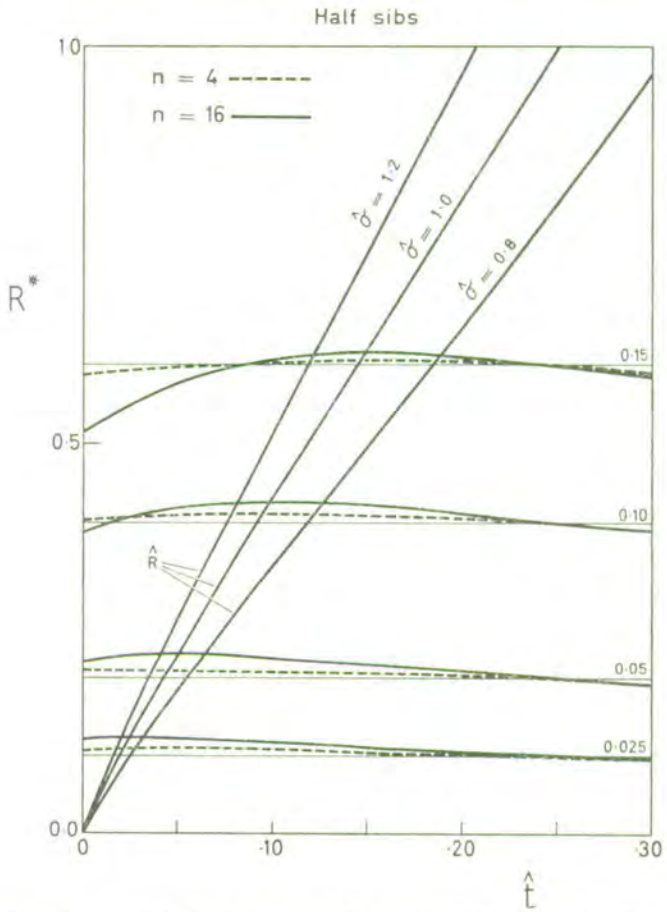

FIG. 1. Achieved response (R^*) plotted against the estimate ($\hat{t}$) of the intra-class correlation (t) for half-sib families of size n and several values of t. The predicted response ($\hat{R}$) is shown for $n = 16$ and three values of the estimate ($\hat{\sigma}$) of the phenotypic standard deviation (σ). For illustration $\sigma = 1$, the heritability equals $4t$ and the horizontal lines are the achieved response from individual selection.

Mean response achieved. If the intra-class correlation and variance are estimated in a design with a total of s families and n progeny in each family, then

$$\hat{\sigma}^2 = [B+(n-1)W]/n, \quad \hat{t} = (B-W)/[B+(n-1)W],$$

where B and W are the mean squares between and within families, respectively in the analysis of variance. Whilst $\hat{\sigma}^2$ is an unbiased estimator of σ^2, there is a small bias, of order $1/s$, in $\hat{t}$ as an estimator of t; this bias turns out not to be important and will be ignored in the subsequent discussion. By numerical integration over the distribution of B and W, $\mathrm{E}(\hat{R})$ and $\mathrm{E}(R^*)$ can be obtained.

When obtaining numerical results, however, it is necessary to decide what to do about unreasonable values of $\hat{t}$, which are values outside the

range 0 to 0·25 for half-sib families assuming no environmental correlation between sibs. The probability that $\hat{t}$ will fall outside this range decreases as the total sample size increases, but is still appreciable even with fairly large samples if t is small (Gill and Jensen, 1968). For example, if t is estimated from 100 sire families each of size 10 there is a 7% chance that $\hat{t}$ will be negative when the true value of t is 0·025.

Again a distinction has been drawn between the full-sib and half-sib situation. With half-sibs it was assumed that the heritability and intra-class

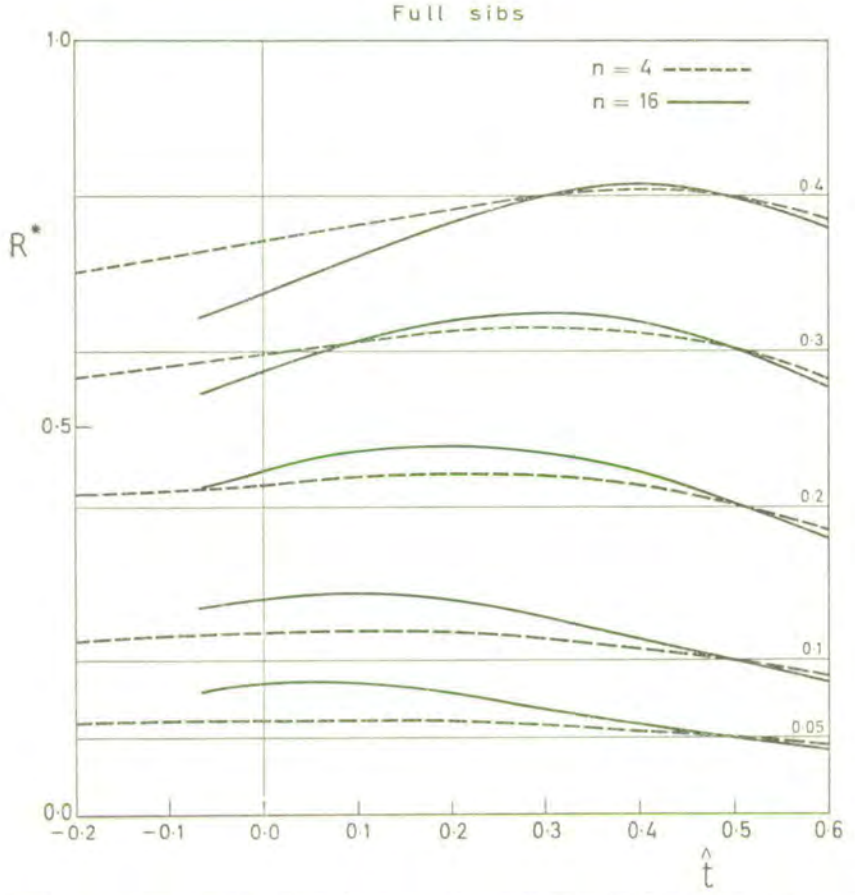

FIG. 2. As Figure 1, but for full-sib families and heritability equal to $2t$, (predicted responses are not shown).

correlation were estimated from the same experiment. The estimate of h^2 (and thus the corresponding estimate of t) was modified if it fell outside the range 0 to 1 by setting $\hat{h}^2$ to the appropriate limiting value. In the full-sib case it was generally assumed that the heritability was estimated elsewhere and the experiment was used to estimate t only. Since a wide range of t values are possible due either to competition between family members or high maternal correlations, no restrictions were put on the values of $\hat{t}$ in deriving the full-sib results.

The expected (i.e. mean) loss in response achieved using estimates of

parameter values relative to that from the optimum index is expressed as a proportion of the optimum response by the 'proportional loss in response', $L = [E(R^*) - R]/R$. Some values of L are given in Table 2(a), obtained by numerical integration. As could be anticipated from Figures 1 and 2, these expected losses are of a few percentage points or less, even with as few as 10 families.

TABLE 2

Proportional loss in efficiency, L = [*E*(R*)−R]/R %, *in an index of individual and family mean performance when the intra-class correlation* (t) *is estimated from* s *family records of the same size* (n). *Values were computed by two methods*

n	*t*	*s*	10	40	160	10	40	160
			Proportional loss (%)					
			(a) Numerical integration			(b) Taylor's series		
Half-sibs			(modified)					
4	0·025		−1·12	−0·41	−0·13	−3·13	−0·78	−0·20
	0·1		−1·25	−0·65	−0·20	−3·29	−0·82	−0·21
16	0·025		−1·35	−0·51	−0·15	−2·41	−0·60	−0·15
	0·1		−2·00	−0·67	−0·17	−2·67	−0·67	−0·17
Full-sibs			(unconditional)					
4	0·05		−1·52	−0·37	−0·09	−1·48	−0·37	−0·09
	0·2		−1·99	−0·52	−0·13	−2·08	−0·52	−0·13
16	0·05		−0·86	−0·21	−0·05	−0·85	−0·21	−0·05
	0·2		−1·65	−0·43	−0·11	−1·74	−0·44	−0·11

More general results can be obtained using the Taylor's series expansion. From (12), assuming $\hat{h}^2 > 0$, it can be shown that

$$\tfrac{1}{2}\frac{\partial^2 R^*}{\partial \hat{t}^2}\bigg|_{\hat{t}=} = -R\,\frac{(n-1)(1-r)^2[1+(n-1)r]^2}{2(1-t)[1+(n-1)t]\{(1-t)[1+(n-1)t]+(n-1)(r-t)^2\}^2}$$

$$= D, \text{ say.} \tag{13}$$

The actual loss depends on the variance of the estimate $\hat{t}$. For an experiment with s families each of size n, this is given by

$$V(\hat{t}) = \frac{2(1-t)^2[1+(n-1)t]^2}{(s-1)n(n-1)} \tag{14}$$

(Fisher, 1925). Thus assuming s is sufficiently large that terms in $1/s^2$ can be ignored, the proportional loss in efficiency is given approximately by

$$L = [E(R^*-R)]/R = DV(\hat{t})/R$$

$$= -\frac{(1-t)[1+(n-1)t](1-r)^2[1+(n-1)r]^2}{sn\{(1-t)[1+(n-1)t]+(n-1)(r-t)^2\}^2} \tag{15}$$

using (7), (13) and (14).

Some examples using (15) are given in Table 2(b) for comparison with the exact results from numerical integration. With half-sibs, predictions of loss using (15) are of the right order of magnitude with as few as 10 sires, and are very good for samples of 160 sires. In the full-sib case, with unconditional

expectations, the fit is uniformly good. Of course, as the number of sires increases, the results from the two methods converge.

Values of D, the coefficient of $V(\hat{\imath})$ in the formula for expected response (*see* Equation 13), and of the proportional loss in efficiency (L) when the same design is adopted for estimation and use of t are given in Figure 3. The latter is expressed in terms of the number of families, so the actual loss is that shown in the graph, divided by the number of families. As an example with full-sib families of size 4 and $t = 0{\cdot}2$ the top right graph gives $D = -0{\cdot}7$ approximately, implying $L = -0{\cdot}7V(\hat{\imath})/R$. The lower right graph gives $Ls = -0{\cdot}21$, equivalent to a proportional loss of $-0{\cdot}0021$ or $-0{\cdot}21\,\%$ from

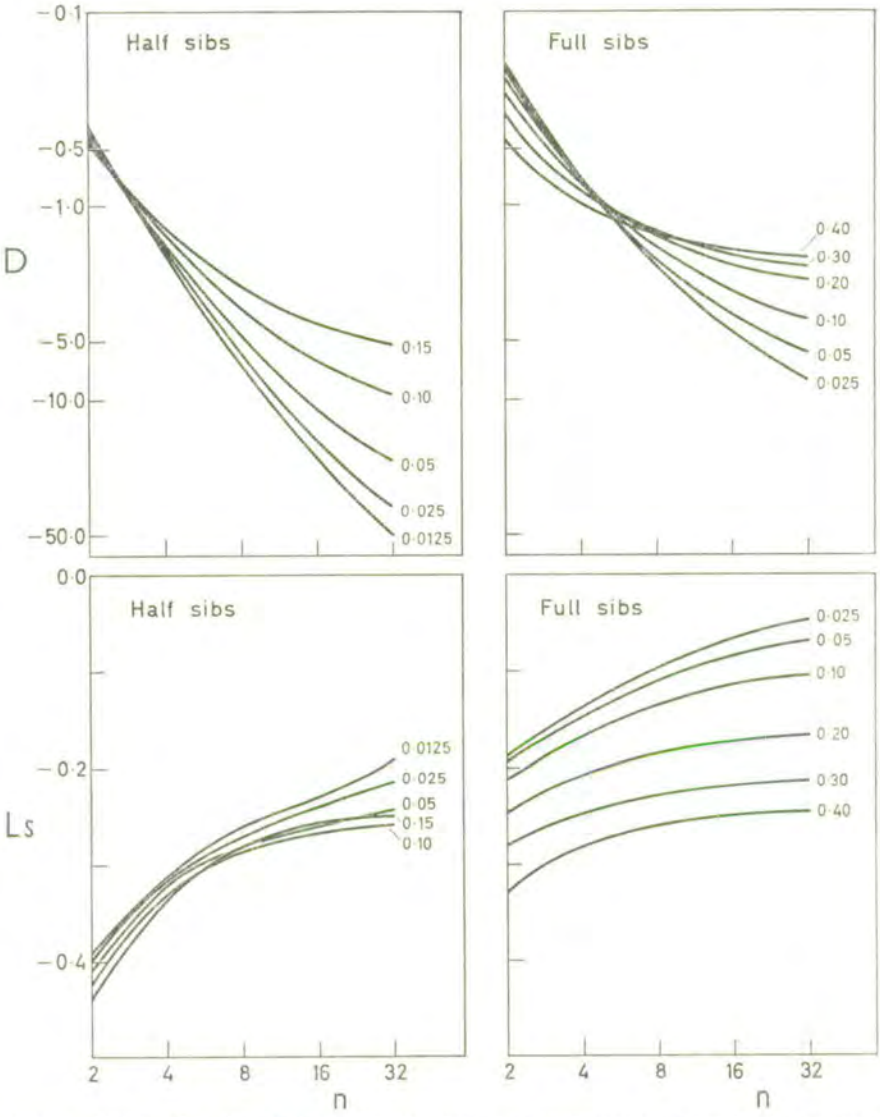

FIG. 3. Values of D, the coefficient of $V(\hat{\imath})$, obtained by differentiation and Ls, the expected proportional loss in response, for several values of the intra-class correlation (t) and full- and half-sib family sizes (n).

an analysis on 100 families. The coefficient, D, is very sensitive to family size: as shown by Figures 1 and 2 the index contributes more to progress with larger families and the curves of $R^\star$ against $\hat{\imath}$ show a more pronounced maximum. The proportional loss is much less sensitive to change in family size, the index sensitivity and accuracy of estimation of t partly compensating for each other.

Mean predicted response. With values of $\hat{\imath}$ modified to lie in the acceptable range $0 \leqslant \hat{\imath} \leqslant 0{\cdot}25$ for half-sibs, the Taylor's series approximation for the mean of the predicted response was found to be of insufficient accuracy and only numerical integration results are given. Some are shown in Figure 4, where values of $E(\hat{R})$ are plotted against t for different numbers of half-sib families used in estimating the parameters. Since $E(R^\star)$ is usually very close to the optimum response, the bias between predicted and achieved response, $E(\hat{R}-R^\star)$, is approximately equal to $E(\hat{R})-R$. It is seen that for small values of the parameter t, progress is overestimated and for large values it is

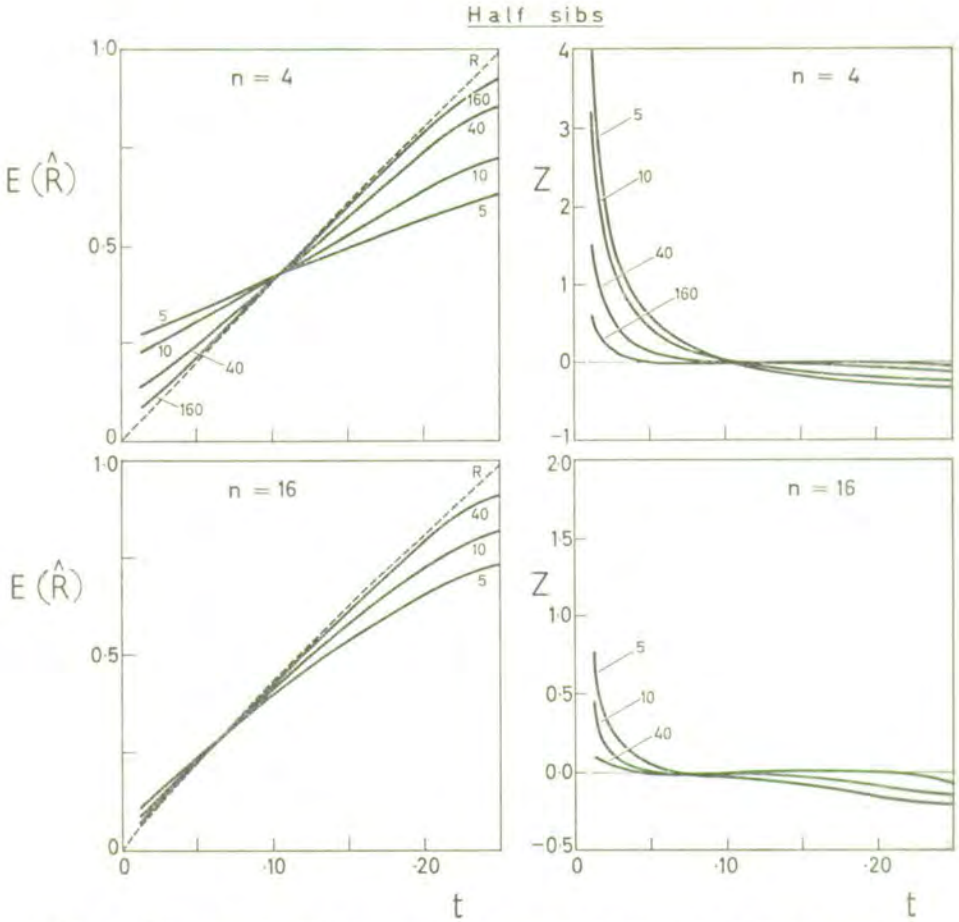

FIG. 4. The mean predicted response, $E(\hat{R})$, and bias in prediction expressed relative to the optimum response, $Z = [E(\hat{R})-E(R^\star)]/R$. Parameters are estimated from s half-sib families of size n for different values of the intra-class correlation (t) and a phenotypic standard deviation of $\sigma = 1$. Results were modified to remove unreasonable values by setting $\hat{R} = 0$ if $\hat{\imath} \leqslant 0$, and $\hat{R} = \hat{\sigma}$ if $\hat{\imath} \geqslant 0{\cdot}25$.

underestimated. The bias can be appreciable: for example, with $s = 40$, $n = 4$ and $t = 0{\cdot}025$ the bias is about 56% of the optimum progress, as shown on the right-hand part of Figure 4 where the bias is expressed relative to the optimum response, i.e. $\mathrm{E}(\hat{R} - R^{\star})/R$.

Individual v. index selection. As illustrated by Figure 1, part of the error in predicting response is contributed by errors in predicting the phenotypic variance and the heritability *per se*. These are not sources of error when alternative selection procedures are being compared, for example individual and index selection. Let R_1, $\hat{R}_1$ and $R^{\star}_1$ denote the optimum, predicted and achieved responses to individual selection. The ratio $\hat{R}/\hat{R}_1$ of predicted responses depends only on the estimate of the intra-class correlation, and from (11) is

$$\hat{R}/\hat{R}_1 = \left\{1 + \frac{(n-1)(r-\hat{t})^2}{(1-\hat{t})[1+(n-1)\hat{t}]}\right\}^{\frac{1}{2}} \tag{16}$$

This quantity exceeds unity, except where $\hat{t} = r$ (a heritability estimate of unity) and is continuous over the relevant range (except at $\hat{h}^2 = 0$ when it is not defined). Values of the predicted ratio are compared in Figure 5 with the ratio achieved $R^{\star}/R^{\star}_1$ ($= R^{\star}/R_1$ since no index weights are needed for individual selection) for a range of values of true parameters and estimates. The results are essentially transformations of those given in Figures 1 and 2. The advantages of using the index tend to be much overestimated if $\hat{t}$ is smaller than t, but slightly underestimated when $\hat{t}$ is larger than t. For example, with full-sib families and $t = 0{\cdot}2$, $R/R_1 = 1{\cdot}10$, whereas with $\hat{t} = 0{\cdot}1$, $\hat{R}/\hat{R}_1 = 1{\cdot}19$ and with $\hat{t} = 0{\cdot}3$, $\hat{R}/\hat{R}_1 = 1{\cdot}05$. Over this range of $\hat{t}$, $R^{\star}/R_1$ lies between 1·09 and 1·10.

Some results for the mean relative advantage of index selection to individual selection are given in Table 3, as follows: R/R_1: the ratio when the true parameter values are used;

$$s[\mathrm{E}(R^{\star}/R_1) - R/R_1] = s[\mathrm{E}(R^{\star}) - R]/R_1:$$

the average difference between the relative advantage when estimates of the parameters are used and when the true parameter values are used, calculated on a single sire basis (the actual difference being $1/s$ of that shown);

$$s\mathrm{E}[\hat{R}/\hat{R}_1 - R^{\star}/R^{\star}_1]:$$

the expected bias between the predicted and realized advantage, again calculated on a single sire basis. The Table shows that the actual bias is about $1/s$ for a wide choice of parameters.

(ii) *Individual, full- and half-sib performance*

Both full- and half-sib information can be included in an index, as originally shown by Osborne (1957). The variables in the index are

x_1 = deviation of individual observation from full-sib family mean,
x_2 = deviation of full-sib family mean from half-sib family mean,
x_3 = half-sib family mean.

Throughout (see Table 1 for definitions) there are assumed to be n individuals in each full-sib family and d full-sib families in each half-sib family (i.e. d dams/sire) so there are nd individuals in each half-sib family. The intra-

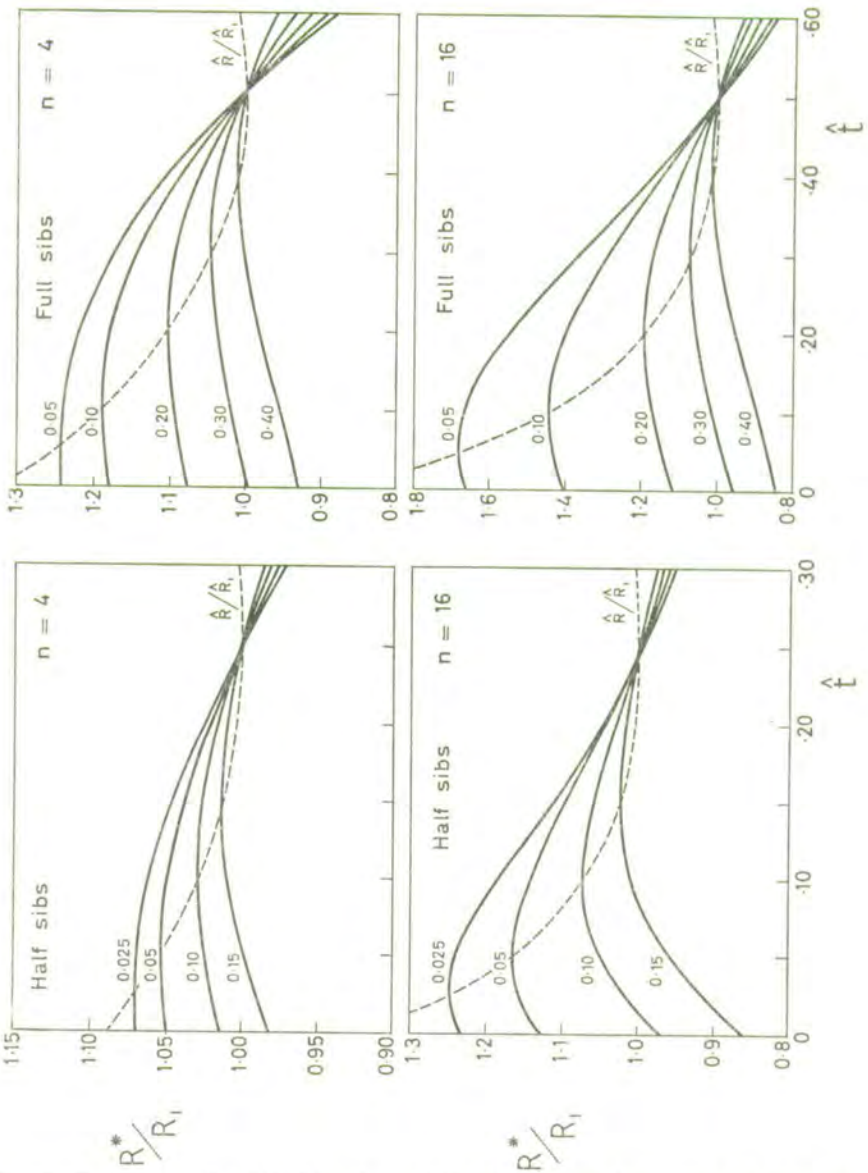

FIG. 5. Ratios of predicted ($\hat{R}/\hat{R}_1$) and actual (R^*/R_1) response from index and individual selection for different values (t) and estimates ($\hat{t}$) of the intra-class correlation and family size (n).

class correlation is t_s between half-sibs and t_d between full-sibs within half-sib families expressed as a proportion of the phenotypic variance. Thus the correlation of full-sibs is t_s+t_d. The relevant matrices are

$$\boldsymbol{P}=\frac{\sigma^2}{nd}\begin{pmatrix} d(n-1)(1-t_s-t_d) & 0 & 0 \\ 0 & (d-1)[1-t_s+(n-1)t_d] & 0 \\ 0 & 0 & 1+(n-1)t_d+(nd-1)t_s \end{pmatrix}$$

and

$$G = \frac{h^2\sigma^2}{4nd}\begin{pmatrix} 2d(n-1) \\ (d-1)(n+2) \\ 2+d+nd \end{pmatrix}.$$

Since h^2 factors out of G, the relative weights given to individual, full- and half-sib means do not depend on heritability.

An investigation was made of the effect of errors in t_s and t_d on $R^\star$. Because so many parameter and estimate combinations are possible, details will not be given. It is found that the efficiency of the index is more sensitive

TABLE 3

Predictions and realizations of the ratio of response using an index of individual and family information to that using individual selection. The intra-class correlation (t) *is estimated from* s *families each of the same size* (n) *as used subsequently. The ratio using the parameter values is* R/R_1 *and the loss* $(E(R^\star/R^\star_1 - R/R_1) = L \times R/R_1)$ *is the difference between the ratio using the optimum and computed index, and the bias* $(E(\hat{R}/\hat{R}_1 - R^\star/R^\star_1)$ *is the difference between the predicted and actual ratio*

n	t	R/R_1	Loss $\times s$	Bias $\times s$
Half sibs				
4	0·025	1·070	−0·335	0·813
	0·1	1·028	−0·338	0·755
16	0·025	1·252	−0·302	0·946
	0·1	1·072	−0·286	0·704
Full-sibs				
4	0·05	1·247	−0·184	0·833
	0·2	1·100	−0·229	0·729
16	0·05	1·681	−0·143	1·221
	0·2	1·192	−0·208	0·828

to poor estimates of t_s than t_d; but, as for single classifications, the index is very robust to wide departures of estimates from parameter values. For example, with $t_s = 0{\cdot}05$ and $t_d = 0{\cdot}1$ corresponding to a trait of heritability 0·2 with some maternal effects or dominance, and $n = d = 4$, a typical situation for pigs, the value of R using the optimum index is 0·2433 in standardized units. With

$\hat{t}_d = 0{\cdot}1$ and $\hat{t}_s = 0{\cdot}0$ 0·1 0·2
$R^\star = 0{\cdot}2374$, 0·2395 and 0·2215, respectively
and with $\hat{t}_s = 0{\cdot}05$ and $\hat{t}_d = 0{\cdot}0$ 0·05 0·2 0·3
$R^\star = 0{\cdot}2419$, 0·2430, 0·2418 and 0·2396, respectively.

When both $\hat{t}_s$ and $\hat{t}_d$ depart from their parameter values, the reduction in efficiency is roughly the sum of the losses caused by the two taken separately. In most hierarchical analyses of variance used to estimate t_s and t_d there are many more degrees of freedom between dams within sires than between sires. Thus the sampling variance of $\hat{t}_s$ is usually much larger than that of $\hat{t}_d$ which, coupled with the greater sensitivity of the efficiency of the index to $\hat{t}_s$ than $\hat{t}_d$, implies that most errors in using an index will come from poor estimates of t_s, the intra-class correlation between sires.

Average values of proportional loss in response, $[E(R^\star) - R]/R$ for estimates from an analysis of variance with s sire families, and with d dams/sire

and n progeny per dam as used subsequently, have been computed using the Taylor's series approximation and numerical differentiation. Some examples are given in Table 4. In each of these examples the proportional loss $\times s$ lies in the range $-0{\cdot}23$ to $-0{\cdot}40$, i.e. the proportional loss is around $-0{\cdot}3$, divided by the number of sire families used in the initial analysis. These values correspond very closely to those shown in Figure 3 for an index using only individual and half-sib family information. Thus for design purposes it is of little concern whether the full-sib family information is collected and used.

TABLE 4

Expected proportional loss in efficiency, L, *for estimates based on* s *families for an index of individual, full- and half-sib family information with intra-class correlation of half-sibs* (t_s) *and full-sibs within half-sibs* (t_d) *estimated and used in families with* d *dams/sire and* n *progeny/dam.*

t_s	t_d	d	4	4	8	2	8
		n	2	4	2	8	8
					Loss $\times s$		
0·025	0·025		−0·313	−0·279	−0·260	−0·321	−0·235
0·025	0·05		−0·314	−0·280	−0·261	−0·321	−0·231
0·05	0·05		−0·329	−0·304	−0·281	−0·358	−0·269
0·05	0·1		−0·331	−0·304	−0·282	−0·353	−0·263
0·1	0·1		−0·351	−0·328	−0·282	−0·397	−0·283

No analysis of errors of predictions of the response $\hat{R}$ has been undertaken. It is clear from Figure 1, however, that the main source of error will be the heritability estimate (i.e. $4\hat{t}_s$) rather than the index weights.

DISCUSSION

It is perhaps surprising that so little attempt has been made previously to consider the effects of errors of parameters on the efficiency of selection indices using individual and family information on one trait. Previous studies, for example those of Harris (1963, 1964), the Pig Industry Development Authority (1965) and Mao (1971), have been primarily concerned with the use of the index for multiple trait selection. Lush (1947, p. 366) remarked, however, 'that the values used for r and t may not be quite correct. Selecting on a combination of family and individuality will often be a little less superior . . . than has been indicated here. This discrepancy will be small, since the correlation between P [individual phenotype] and $\bar{Y}$ [mean phenotype of family] will cause whichever one of them is overemphasized to pick up part of the load which should have been carried by the other'. Our results indicate that Lush's intuition was correct, and perhaps others were wise enough to believe it and undertake no check on the assumptions.

To illustrate some of the results derived in the paper, consider selection in pigs using full-sib families of size 4. It can be seen from Figure 3 that for $n = 4$ the proportional loss times the number of families used to estimate t lies between about 15% and 30% for values of the intra-class correlation between 0·025 and 0·4. Therefore it will need an initial experiment of 30 full-sib families of size 4 to reduce the expected loss to 1%, essentially regardless of the true value of the intra-class correlation. The fact that the actual

losses are very small is no surprise if one looks at Figure 2. The curves for R^* are generally fairly flat. This indicates that even if the estimate of t is a long way from the true value very little loss of progress will result. Figure 2 also shows that unless the true intra-class correlation is very high it is generally better to guess a reasonable value of t, if no data is available, and construct an index using that rather than use individual selection.

In an experimental comparison between selection on individual performance, family mean and an index of both, Wilson (1974) did not obtain a benefit from using the index; indeed greater responses were made from individual selection. It is clear from our results that a poor estimate of the intra-class correlation of sibs is not a sufficient explanation and that others must be sought.

This decision of whether to include information from relatives in an index or simply select individuals on their phenotype is often made from a comparison of the predicted responses from index and individual selection. Table 3 shows that the real benefits from using index selection tend to be overestimated but the bias is generally small. With full sibs, $n = 4$, $t = 0{\cdot}05$, the best index would be 1·247 times as efficient as individual selection, giving a gain of 24·7%. With an estimate of t obtained from 30 families the gain on average would be $-0{\cdot}184 \times 100/30\%$ less than this, that is only approximately 24·1%, whereas the predicted gain on average would be 27·5%. However, Figure 5 shows that unless $\hat{t}$ is close to the true value of t, the ratio $\hat{R}/\hat{R}_1$ may give a poor estimate of the true advantage in using the index, R^*/R^*_1. The probability of $\hat{t}$ being close to the true value depends on the sampling distribution of $\hat{t}$ in the initial experiment. For example if $t = 0{\cdot}2$ and 30 full sib families of size 4 are used in the experiment then the standard error of $\hat{t}$ is about 0·1 and it can be seen that the ratio $\hat{R}/\hat{R}_1$ may give little indication of the actual advantage.

If only the progress achieved is of importance then the estimate of the intra-class correlation can have a high sampling variance without much loss in mean response. For example, with half-sib families of size 16, typical of beef cattle in a testing regime, the proportional loss in efficiency, L, is in the range $-16\,V(\hat{t})$ to $-6\,V(\hat{t})$ (Figure 3) for heritabilities in the range 0·1 to 0·4; when s families are used to estimate t the loss is about $25/s\%$. So to get an expected loss of 1% or less about 25 families of size 16 are needed. If the intra-class correlation is really 0·1 (and the heritability 0·4) then with 50 sires the loss is reduced to 0·5%, which is more or less negligible in practical terms, whereas the standard error of $\hat{t}$ is still about 0·03. The estimate of progress, $\hat{R}$, depends critically on what value of $\hat{h}^2$ is used and the standard error of $\hat{R}$ has been shown to be similar in magnitude to the standard error of $\hat{h}^2$. In the previous example, the standard error of $\hat{R}$ (obtained from simulation) was shown to be about 0·12; the optimum progress in this case is 0·43. Thus if it is important to make accurate predictions of the amount of progress that is likely to be made, a much larger initial experiment is required than if one only wishes to be fairly certain that the selection index will be reasonably efficient.

Throughout the paper it has been assumed that the initial experiment was completely balanced and that the family size used in the experiment was also used subsequently to evaluate individuals. These assumptions may break down in several ways in practice. It has been shown using the approximate results, however, that the expected loss is the product of a constant, D, and

the variance of the intra-class correlation coefficient. The value of D depends on the family structure used in the index (as shown in Figure 3) and the variance of $\hat{t}$ depends on the size and structure of the initial experiment. Consequently the expected loss may readily be calculated in those cases where the family size used in the index is different from that used in the experiment. It also enables the expected loss to be calculated if t is estimated from an unbalanced experiment or from regression as long as the sampling variance of $\hat{t}$ is known. It may also happen that information is available on different numbers of relatives when the index is used in practice. Henderson (1963) has discussed this in the case where the index weights are known. When the index weights are estimated the additional loss in efficiency will depend on the distribution of family size amongst the individuals. If individuals are selected on their index score, regardless of how many relatives were measured, then it can be shown that the expected loss is a weighted mean of the expected losses for each family size. The magnitude of D increases with family size and so the expected loss also increases with family size for a given initial amount of data. Therefore the expected loss, when individuals are selected with different amounts of information, will be less than the expected loss calculated with the largest family size present.

If selection is practised in a population the genetic parameters subsequently change. With genes of small effect there would be a reduction in heritability and correlation among sibs (e.g. Bulmer, 1971), but the change cannot be predicted more generally without knowledge of gene frequencies. Whilst some change in t could be allowed for in calculating the selection index, in view of the robustness of the achieved response to errors in $\hat{t}$ (Figures 1, 2 and 5) it is probably unnecessary. Of course, if the parameters are estimated in the population after some selection is practised they will be more precise.

In order to compute expected losses in efficiency and errors in predictions from estimates of parameters based on samples of data it has been assumed that no prior information is available on the population or, in the literature, on other populations likely to be similar. It is not known how to incorporate such information, but clearly this ought to be attempted. Formally, the new estimate could be regressed towards other values but no weights can be given. To take a more extreme view, it is clear from Figure 5, for example, that if a value of about 0·1 is used tor $\hat{t}$ with half-sib families and 0·3 with full-sib families, the index will give a response which does not differ much from the optimum, regardless of the true value of t.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the Meat and Livestock Commission. We are grateful to Marjorie McEwan for computational assistance.

REFERENCES

BULMER, M. G. 1971. The effect of selection on genetic variability. *Am. Nat.* **105**: 201–211.

FISHER, R. A. 1925. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

GILL, J. L. and JENSEN, E. L. 1968. Probability of obtaining negative estimates of heritability. *Biometrics* **24**: 517–526.

HARRIS, D. L. 1963. The influence of errors of parameter estimation upon index selection. In *Statistical Genetics and Plant Breeding* (ed. W. D. Hanson and H. F. Robinson),

pp. 491–500. Publ. 982. National Academy of Sciences—National Research Council, Washington, D.C.

HARRIS, D. L. 1964. Expected and predicted progress from index selection involving estimates of population parameters. *Biometrics* **20**: 46–72.

HAZEL, L. N. 1943. The genetic basis for constructing selection indexes. *Genetics, Princeton* **28**: 476–490.

HENDERSON, C. R. 1963. Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (ed. W. D. Hanson and H. F. Robinson), pp. 141–163. Publ. 982. National Academy of Sciences—National Research Council, Washington, D.C.

HILL, W. G. 1974. Heritabilities: estimation problems and the present state of information. In *1st Wld Cong. Genet. appl. Livestock Prod.* **1**: 343–351. Editorial Garsi, Madrid.

LUSH, J. L. 1947. Family merit and individual merit as bases for selection. *Am. Nat.* **81**: 241–261; 362–379.

MAO, I. L. 1971. The effect of parameter estimation errors on the efficiency of index selection and on the accuracy of genetic gain prediction. *Ph.D. Thesis, Cornell University, Ithaca, New York.*

OSBORNE, R. 1957. The use of sire and dam family averages in increasing the efficiency of selective breeding under a hierarchical mating system. *Heredity, Lond.* **11**: 93–116.

PIG INDUSTRY DEVELOPMENT AUTHORITY. 1965. *Combined Testing. Recommendations by the Statistics Section for the Selection Index*, (Mimeograph). DA 188. Pig Industry Development Authority, London.

WILSON, S. P. 1974. An experimental comparison of individual, family and combination selection. *Genetics, Princeton* **76**: 823–836.

(*Received* 12 *June* 1975)

APPENDIX

Approximations using Taylor's series. The quantity R^*, for example, is a function of the estimates $\hat{G}$ and $\hat{P}$ of the parameters G and P. The elements of these matrices can be represented by the vector of estimates $\hat{y}$ of parameters y. More simply $\hat{y}$ and y can be reduced solely to the minimum number of parameters needed (e.g. the intra-class correlation for an index based on individual and sib performance). Using Taylor's series to express $R^*(\hat{y})$ about $R^*(y)$, the value using the parameters themselves,

$$R^*(\hat{y}) = R^*(y)+\Sigma(\hat{y}_i-y_i)\frac{\partial R^*}{\partial \hat{y}_i}\bigg|_{\hat{y}=y} +\tfrac{1}{2}\Sigma\Sigma(\hat{y}_i-y_i)(\hat{y}_j-y_j)\frac{\partial^2 R^*}{\partial \hat{y}_i \partial \hat{y}_j}\bigg|_{\hat{y}=y}+\ldots \qquad \text{(A1)}$$

Now, $R^*(y) = R$, the response from the optimum index, and at the optimum $\dfrac{\partial R^*}{\partial \hat{y}_i} = 0$ for all i. If the parameter estimates are unbiased then by taking expectations over equation (A1)

$$\mathrm{E}(R^*) = R+\tfrac{1}{2}\Sigma\Sigma\,\mathrm{Cov}\,(\hat{y}_i, \hat{y}_j)\frac{\partial^2 R^*}{\partial \hat{y}_i \partial \hat{y}_j}\bigg|_{\hat{y}=y}, \qquad \text{(A2)}$$

plus higher order terms, assumed to be small. The method used by Harris (1963, 1964) was similar, but less direct.

22

Effect of sampling errors on efficiency in selection indices II. Use of information on associated traits for improvement of a single important trait

by

Jill Sales and William G. Hill

Anim. Prod. 23(1) (in press)

EFFECT OF SAMPLING ERRORS ON EFFICIENCY OF SELECTION INDICES

II. USE OF INFORMATION ON ASSOCIATED TRAITS FOR IMPROVEMENT OF A SINGLE IMPORTANT TRAIT

Jill Sales and William G. Hill
Institute of Animal Genetics, Edinburgh EH9 3JN

SUMMARY

An analysis is undertaken of the effect of errors in estimates of parameters on the response to selection for an economically important trait (trait 1) when one or more additional traits are added in a selection index. The detailed analysis is confined to one additional trait (trait 2) which contributes useful information unless the genetic and phenotypic regressions of trait 1 on trait 2 are equal.

If there are errors in parameter estimates the extra response obtained by including trait 2 will usually be over-predicted. When trait 2 actually contributes no useful information the predicted benefit equals the real loss in efficiency from its inclusion.

The loss in efficiency from poor estimation of parameters, whether or not the second trait makes a contribution, is roughly one-quarter of the squared coefficient of variation of a heritability estimate of trait 1 in the same experiment.

INTRODUCTION

Selection indices are often used in improvement schemes where information is available on several traits on each animal. The economic merit of an animal may depend on some of these traits, and others may be incorporated into an index only to improve the accuracy of selection. These additional traits may be quantitative measurements or they may be the genotype for some blood group or biochemical variant.

The selection index is a linear combination of the observed measurements constructed so as to maximise the correlation with breeding value and thus response in economic merit. This maximisation will only be realised if the underlying genetic and phenotypic parameters are known exactly. In practice these parameters have to be estimated from samples of data and use of the estimates rather than the true parameters will lead to errors in predicting the response from the index and to a loss of efficiency relative to using the optimum index. Such errors arising from single trait selection, when the index comprised individual and family records, were discussed in Part I of this series (Sales and Hill, 1976). This paper considers the use of an index of several traits for maximising response on a single trait. For simplicity and purposes of illustration most of the results are concerned with a two trait index. Several analyses of the effect of errors in multiple trait selection have already been published and these are listed in Part I. An analysis of the potential benefits from incorporating additional quantitative traits in an index has been made by Gjedrem (1967), but without considering the effects of errors in parameter estimates.

Neimann-Sörensen and Robertson (1961) and Smith (1967) have discussed the incorporation of blood group or biochemical markers in a selection index.

THEORY AND METHODS

Let trait 1 be the trait of economic importance and trait 2 be of no direct importance, although possibly of use in increasing the accuracy of selection for trait 1. Let $\underset{\sim}{P}$ and $\underset{\sim}{G}$ be the phenotypic and genetic covariance matrices between traits 1 and 2 (see Part I for definitions), with

$$\underset{\sim}{P} = \begin{pmatrix} \sigma_1^2 & r_P\sigma_1\sigma_2 \\ r_P\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad , \quad \underset{\sim}{G} = \begin{pmatrix} h_1^2\sigma_1^2 & \\ r_G h_1 h_2 \sigma_1\sigma_2 & \end{pmatrix}$$

where σ_i^2, h_i^2 are the phenotypic variances and heritabilities of trait i, and r_P, r_G and r_E are the phenotypic, (additive) genetic and environmental correlations between the traits with $r_E = (r_P - r_G h_1 h_2)/\sqrt{(1-h_1^2)(1-h_2^2)}$. (If there are more traits of no direct importance, the dimensions of $\underset{\sim}{P}$ and $\underset{\sim}{G}$ are increased accordingly). In the most efficient index, the weights are given by the product $\underset{\sim}{P}^{-1}\underset{\sim}{G}$, with

$$b_1 = \frac{(h_1^2 - r_G r_P h_1 h_2)}{1-r_P^2} \quad , \quad b_2 = \frac{(r_G h_1 h_2 - r_P h_1^2)\sigma_1}{(1-r_P^2)\sigma_2} \qquad (1)$$

where b_1 and b_2 are the weights given to traits 1 and 2 respectively (Gjedrem, 1967).

It is clear that if the two traits have a positive genetic correlation, then relatively more weight will be given to the second trait if the traits have a negative phenotypic correlation and vice versa. If $r_G h_2 = r_P h_1$, i.e. when the genetic and phenotypic regressions of trait 2 on trait 1 are equal, the second trait makes no useful contribution, and should be ignored in an efficient index; this, of course, includes the case of two uncorrelated traits ($r_G = r_P = 0$).

The optimum response, R, is given by $R = (\underset{\sim}{G}'\underset{\sim}{P}^{-1}\underset{\sim}{G})^{\frac{1}{2}}$, where, for simplicity, the selection differential is assumed to equal one standard deviation. If the signs of both r_G and r_P are changed, there is no change in the response; for example, a population with $r_G = 0.2$ and $r_P = 0.1$ is equivalent to one with $r_G = -0.2$ and $r_P = -0.1$.

If the index is computed from estimates of the parameters, more quantities need to be defined, for details see Sales and Hill (1976). Let:

$\hat{R}$ = predicted response using the estimates of parameters $\hat{\sigma}_i^2$, $\hat{h}_i^2$, $\hat{r}_P$, $\hat{r}_G$ and the index weights $\hat{b}_i$ computed from these estimates; and

R^* = response actually achieved using the estimated weights $\hat{b}_i$ in a situation where the true parameter values apply.

If selection is practised only on trait 1, the response is $R_1 = h_1^2\sigma_1$. The predicted response is $\hat{R}_1 = \hat{h}_1^2\hat{\sigma}_1$ and the response achieved, R_1^*, also equals R_1 (provided individuals are ranked positively on their phenotype for trait 1).

Values of $\hat{R}$ and R^* can be computed for any specific set of parameter estimates. However, it is useful to know what will be

the expected or mean values of predicted and achieved responses from using indices computed from repeated samples of data. In particular, let $L = (E(R^*)-R)/R$ be the expected proportional loss in response and $Z = E(\hat{R})-E(R^*)$ be the bias in predicted relative to achieved response.

Each quantitative trait is assumed to be normally distributed and the parameters estimated from the mean squares and crossproducts of an analysis of variance of s half-sib families of equal size, n. Most of the results were derived using a Taylor's series approximation (see Appendix to Part I), with the differentiation carried out numerically if direct analysis was too complicated. In the Taylor's approximation both L and Z are inversely proportional to the number of families (s). Appropriate checks of the approximation were made by Monte Carlo simulation. Sets of data were generated by simulating mean squares and crossproducts directly (Hartley and Harris, 1963), rather than via an analysis of simulated normal deviates. Both in practice and with simulated data some parameter estimates may lie outside their permitted bounds ($0 \leq h^2 \leq 1$, $-1 \leq r_G, r_p \leq 1$). In practice one is unlikely to select for a trait with a negative heritability estimate, but in an index one would probably set the estimate to zero. Two sets of simulation results were therefore obtained. In the first (unmodified) all estimates were used regardless of whether they were possible values. In the second (modified) estimates outside the bounds were put equal to the corresponding boundary values. These modifications had most effect at low values of heritability and low family sizes. Obviously, as the number of sire families increases the parameters are estimated more accurately

and the two sets of results become equivalent.

RESULTS

Expected responses

Values for the maximum attainable progress (per unit selection differential assuming $\sigma_1^2 = 1$) together with those for the proportional loss, computed using the Taylor's approximation, are shown in Table 1. The value of L for an initial experiment with s half sib families and n progeny per family is calculated from Table 1 by dividing by the appropriate value of T(=ns), the total number recorded. For example, with $h_1^2 = 0.2$, $h_2^2 = 0.5$, $r_G = r_P = 0.0$ and families of size 4 the value in Table 1 is LT = -89.6. Thus, if the initial experiment comprised 250 families of size 4, the expected loss would be 89.6/(4 x 250) = 0.09 or 9%. Over a wide range of parameters the proportional loss, for a particular family size, is seen to depend critically on h_1^2, the heritability of the economically important trait, and to a much lesser extent on h_2^2, r_G and r_E; this is particularly true for higher values of h_1^2 and small families.

For all the populations considered in Table 1, the difference, $E(\hat{R})-R$, between the mean value of the estimate of progress and the optimum progress was positive and of very similar magnitude to the actual loss $E(R^*)-R$ which equals LR. The bias, expressed as a proportion of the response is therefore approximately equal to 2L and the actual bias, Z is 2L x R.

The results in Table 1 were checked using simulation. The approximations using Taylor's series agreed well, generally within

Table 1 Optimum response (R) and the proportional loss (L) x number of progeny recorded (T = sn) in an initial analysis of data from s half-sib families of size n. ($\sigma_1^2 = 1$)

r_G	r_E	r_P	R	LT		r_P	R	LT	
				n=4	n=16			n=4	n=16
		$h_1^2 = 0.2, h_2^2 = 0.2$				$h_1^2 = 0.2, h_2^2 = 0.5$			
0.0	0.0	0.00	0.200	-79.6	-36.8	0.00	0.200	-89.6	-58.9
0.0	0.5	0.40	0.218	-76.0	-32.8	0.32	0.211	-86.8	-54.2
0.5	-0.5	-0.30	0.261	-60.8	-20.3	-0.16	0.277	-58.4	-22.4
0.5	0.0	0.10	0.216	-74.4	-29.0	0.16	0.238	-72.0	-32.6
0.5	0.5	0.50	0.200	-79.6	-36.8	0.47	0.212	-84.4	-49.0
		$h_1^2 = 0.5, h_2^2 = 0.2$				$h_1^2 = 0.5, h_2^2 = 0.5$			
0.0	0.0	0.00	0.500	-13.9	- 8.6	0.00	0.500	-15.4	-13.4
0.0	0.5	0.32	0.527	-13.4	- 7.8	0.05	0.516	-15.1	-12.8
0.5	-0.5	-0.16	0.555	-12.5	- 6.2	0.00	0.559	-13.2	- 8.5
0.5	0.0	0.16	0.506	-13.5	- 7.7	0.25	0.516	-14.5	-10.9
0.5	0.5	0.47	0.508	-13.7	- 8.2	0.50	0.500	-15.4	-13.4

Table 2 Number of observations (T=sn) required to get an increase in expected response using two traits, i.e. $E(R^*)$ $R_1 = h_1^2\sigma_1$, from an initial experiment with s half-sib families of size n.

h_1^2	h_2^2	r_G	r_P	R/R_1	T	
					n=4	n=16
0.2	0.5	0.0	0.10	1.005	17800	11632
		0.0	0.32	1.054	1696	1056
		0.5	-0.16	1.386	208	80
		0.5	0.16	1.187	456	208
		0.5	0.47	1.062	1432	832
0.5	0.5	0.0	0.10	1.005	3080	2672
		0.0	0.25	1.035	476	400
		0.5	0.00	1.118	124	80
		0.5	0.25	1.035	456	336
		0.5	0.50	1.000	-	-

20% of the modified values (see methods) obtained by Monte Carlo simulation, except for experiments having less than 100 sires and the lower values of both the heritability of trait 1 ($h_1^2 = 0.2$) and family size (n = 4). The approximation tends to overestimate the losses, but gives a good indication of the general magnitude of losses likely to be encountered. The simulation results also confirmed that the bias approximately equals twice the loss in efficiency.

Where $r_G h_2 \neq r_P h_1$ inclusion of the second trait in the index always increases efficiency if accurate estimates of parameter values are available; but its inclusion might actually reduce response if these estimates are poor. Thus a "break-even" number of families or size of initial experiment can be computed where $E(R^*) = R_1 = h_1^2 \sigma_1$, i.e. the expected response using the index equals the response obtained from selecting only on the economically important trait. In terms of proportional loss, the absolute value of L must be less than $1-R_1/R$. These results can be obtained by calculation from Table 1, but for illustration some are given in Table 2. These values are approximations, since the actual numbers required will depend on whether results are modified to remove unreasonable estimates, but the results are of the right order of magnitude. As might be expected, the number of observations required increases as the benefit from including the second trait decreases, and if the second trait can contribute very little useful information a very large experiment would be required before, on average, it would be an advantage to include it in an index.

Distribution of ratios of responses

A decision to include the second trait in the index is likely to be based, at least in part, on the relative magnitudes of $\hat{R}$ and $\hat{R}_1$,

i.e. the predicted responses with and without the second trait included. We have analysed the simple ratio ($\hat{R}/\hat{R}_1$) and that achieved ($R^*/R_1^* = R^*/R_1$). For illustration, in a series of simulated experiments with 100 replicates each and s = 50 and n = 16, values of R^*/R_1 are plotted against $\hat{R}/\hat{R}_1$ in Figure 1 for examples in which, with $h_1^2 = 0.2$ and $h_2^2 = 0.5$ and correct parameter values, the second trait contributes (a) no useful information ($r_G = r_P = 0.0$), (b) very little information ($r_G = 0.5$, $r_P = 0.47$) and (c) about 20% extra information ($r_G = 0.5$, $r_P = 0.16$). Because of the skewness and bounds on the distribution ($R^*/R_1^* \leq R/R_1$) the simulation shows that the means of $\hat{R}/\hat{R}_1$ and R^*/R_1^* are not fully adequate descriptions of their properties. Thus, in case (b) for example, there tends to be a large proportion of replicates in which $R^*/R_1 > 1$, i.e. where a benefit would be made from including the second trait, and a small number where R^*/R_1 1. The predicted ratio, $\hat{R}/\hat{R}_1$, is clearly not a very good guide in any individual experiment, and where a very large gain is predicted there is likely to be real loss for the estimates of parameter values are very poor. The problem of decision making will be discussed subsequently.

Inclusion of a worthless trait

When the second trait contributes no useful information ($r_G h_2 = r_P h_1$) errors in estimates are most likely to lead to wrong decisions. A more complete analytical treatment is also possible in this case, particularly when $r_P = r_G = 0$, and is therefore useful for illustration. Thus, in this section, $r_G h_2 = r_P h_1$ throughout.

Unless the estimates of the parameter values are exactly correct some weight would always be given to trait 2, even though it is of no real use. One would then always predict that more progress is

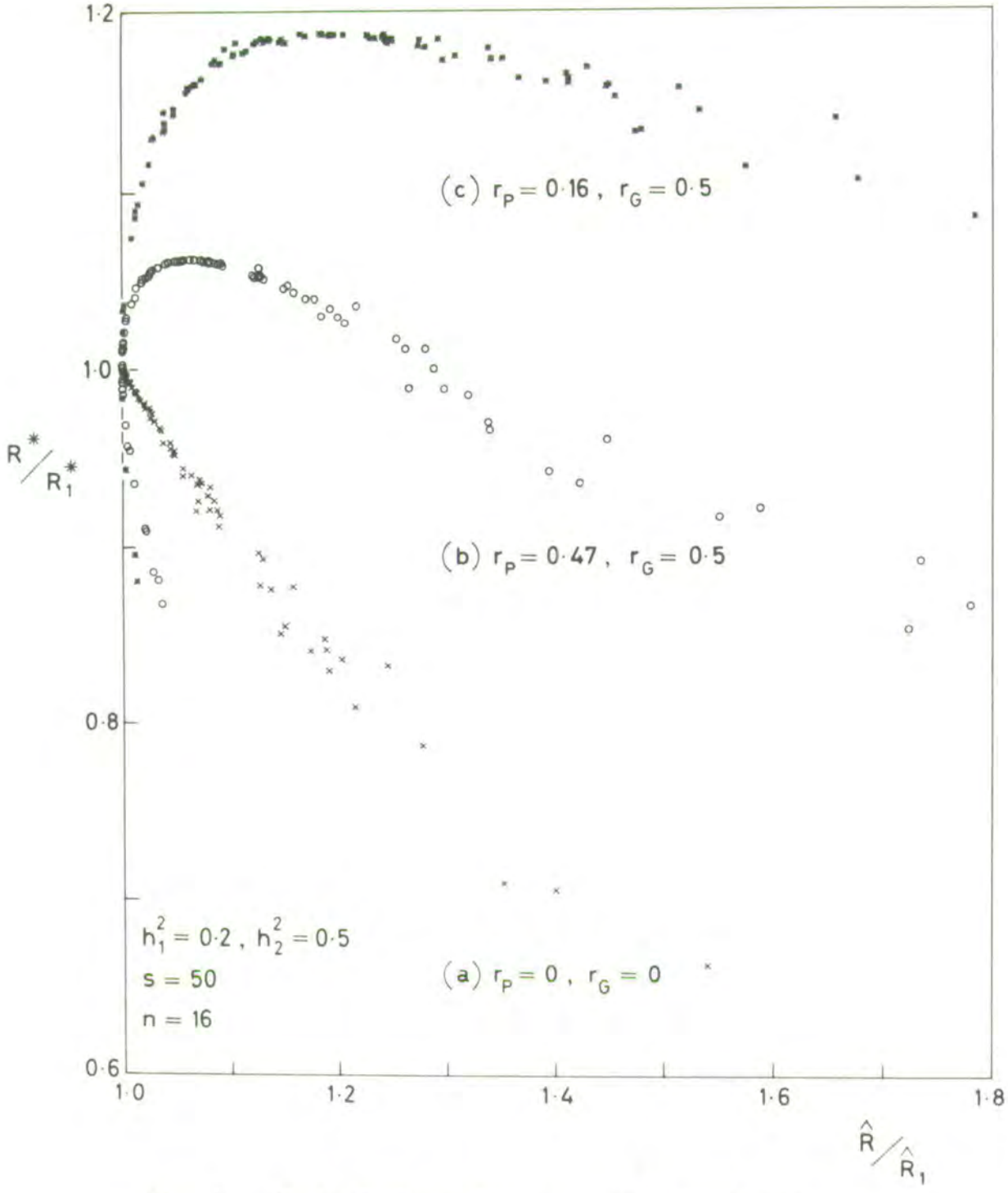

Fig. 1. Monte Carlo replicate populations showing the distribution of ratios of predicted ($\hat{R}/\hat{R}_1$) and achieved (R^*/R^*_1) responses from index and single trait selection.

being made whereas, in fact, less progress is made than if only the economically important trait is used. This is shown in Figure 1a for $r_P = r_G = 0$. Furthermore, Figure 1 demonstrates that least progress is actually made in those situations where most benefit is predicted from including a second trait, i.e. when the estimate of genetic correlation is furthest from its correct value in this case of zero.

The expected values of $\hat{R}/\hat{R}_1$ and R^*/R_1^* may be computed approximately by using a Taylor's series expansion. It can be shown that

$$E(\hat{R}/\hat{R}_1) = 1 + \frac{\sigma_2^2(1-r_P^2)}{2\sigma_1^2 h_1^4} V(\hat{b}_2) \qquad (2)$$

$$= 1 + \lambda,$$

say, and

$$E(R^*/R_1^*) = 1 - \lambda .$$

Thus, on average, the mean predicted relative gain from using the second trait is equal to the real relative loss from its inclusion. When the second trait is contributing no information, use of the first trait alone gives maximum progress and so $R_1^* = R$, and therefore the proportional loss, L, is equal to $-\lambda$. Some particular values are given in Table 1, and a wider set for $r_G = r_P = 0$ in Figure 2, again with the loss expressed as LT where $T = ns$. Table 1 and Figure 2 show that the errors in prediction are greatest when the heritability of the trait of economic importance is low, for then the tendency will always be to give more weight to the other trait if the traits appear to have any useful correlation.

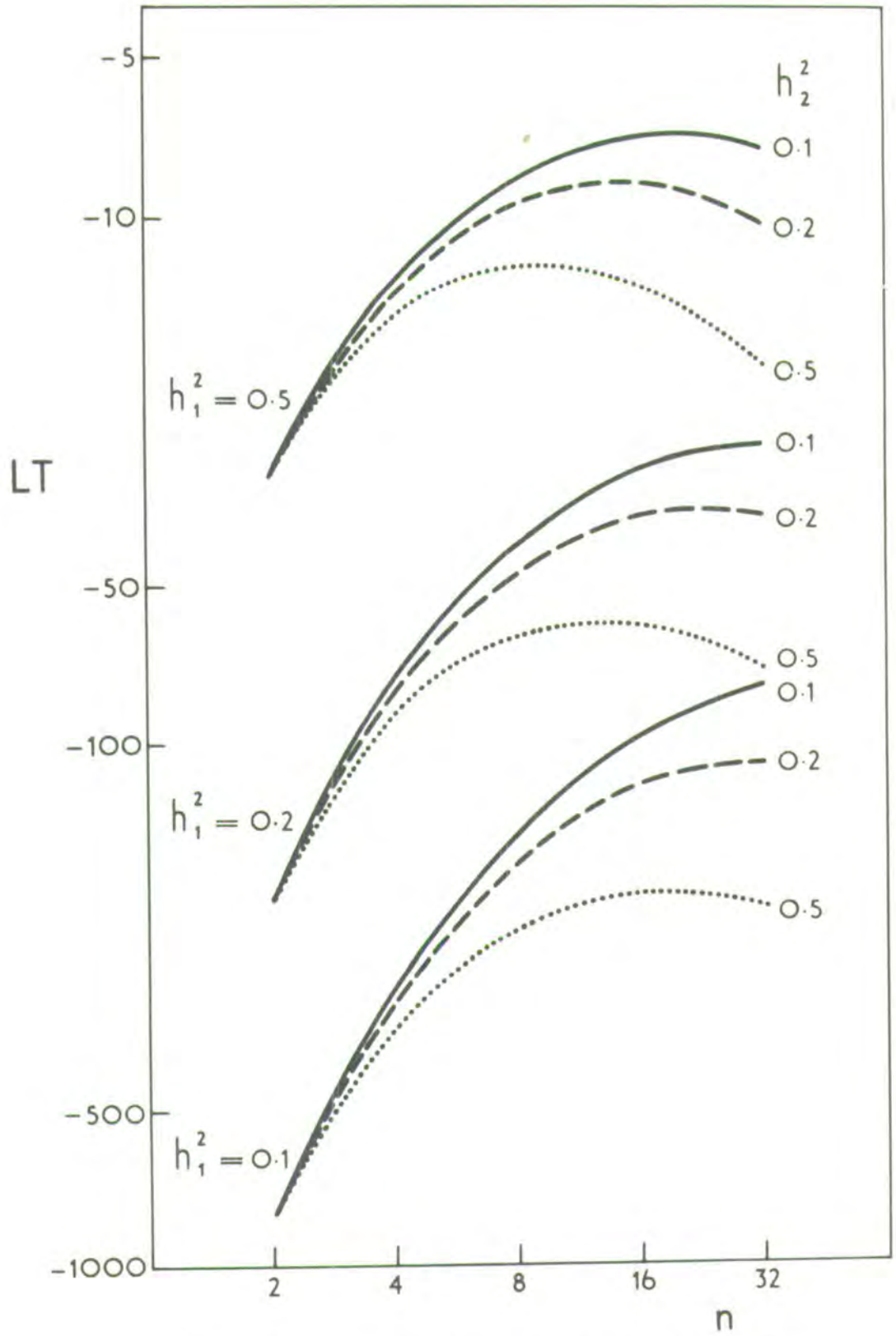

Fig. 2. Expected proportional losses x size of sample (LT) for two uncorrelated traits and a range of heritability values.

Using (2) it can be shown that when estimates are obtained from s half-sib families of size n,

$$\lambda = \frac{(1-t_1)^2[1+(n-1)t_1]^2}{2sn(n-1)t_1^2}\left\{1 + \frac{(t_2-t_1)}{n(1-r_p^2)}\left[\frac{(n-1)^2}{1+(n-1)t_1} - \frac{1}{1-t_1}\right]\right\} \qquad (3)$$

where $t_1 = h_1^2/4$ and $t_2 = h_2^2/4$. Noting that, in the same experiment,

$$V(\hat{t}_1) = \frac{2(1-t_1)^2[1+(n-1)t_1]^2}{sn(n-1)}$$

it can be seen that λ is proportional to $\frac{1}{4}V(\hat{t}_1)/t_1^2$, which is also equal to $\frac{1}{4}V(\hat{h}_1^2)/h_1^4$ (one quarter of the squared coefficient of variation of $\hat{h}_1^2$). When the two traits have equal heritability, the second term in (3) vanishes and $\lambda = \frac{1}{4}V(\hat{h}_1^2)/h_1^4$. When the traits are uncorrelated, (3) can also be written as

$$\lambda = \frac{(1-t_1)[1+(n-1)t_1][1+(n-2)t_2-(n-1)t_1t_2]}{2sn(n-1)t_1^2} \qquad (4)$$

and is evaluated in Figure 2.

For a given total number, T = ns, of animals recorded, it can be shown that if the traits are uncorrelated, the optimum family size for minimising λ is $n = 1+\left[\frac{1-t_2}{t_1(1-t_1)t_2}\right]^{\frac{1}{2}}$. Unless t_1 and t_2 are very different this will not depart far from $n = 1 + 1/t_1$, ($1/t_1$, approximately), obtained by Robertson (1959) as the optimum family size for estimating the heritability of trait 1.

The genetic parameters may, of course, also be estimated from the regression of progeny on parent performance when measurements on both traits are made in each generation. There are many possible

experimental structures, involving full- or half-sib families, measurements on one or both sexes, and possible selection of parents. As a simple example, assume there are s half sib families of size n with performance measured on the unselected sires and their progeny. When the traits are uncorrelated it can be shown using (2) that

$$\lambda = \frac{1}{8nsh_1^4} \left[8 + (h_1^2+h_2^2)(n-1) - 2nh_1^2h_2^2\right] .$$

Since $V(\hat{h}_1^2) = \frac{1}{ns}\left[4 + h_1^2(n-1) - nh_1^4\right]$, it follows that $\lambda = \frac{1}{4}V(\hat{h}_1^2)/h_1^4$ if $h_1^2 = h_2^2$, exactly as in the half-sib analysis of variance case analysed previously.

It may also be possible that information is already available on the phenotypic and genetic parameters of the economically important trait and the new data collection and analysis is undertaken merely to obtain estimates on the second trait and its correlation with the first. It turns out that if the traits are uncorrelated, the expected proportional gains and losses are exactly the same as if there is no prior information on the first trait. Expressed another way, with uncorrelated traits the errors in index computation derive from errors of prediction of the phenotypic and genetic covariances. If the traits are correlated, but trait 2 contributes no useful information because the genetic and phenotypic regressions are equal, it turns out that somewhat more accuracy is lost if only the variances and covariances involving the second trait are estimated in the new experiment.

Inclusion of information on genotypes

A similar situation arises if the genotype of an animal is known at a particular polymorphic locus, for example a blood group or

biochemical variant, and it is thought that these genotypes may influence the trait of economic importance.

For example, assume there are two alleles at the locus with no dominance. The individuals may be typed AA, AB or BB and given a corresponding score, x_2, of 1, 0 or -1. The score may be treated as an additional measurement (trait 2) with unit heritability and incorporated into an index in much the same way as a phenotypic measurement. Then, if q is the frequency of A and $\underline{a}$ is the difference in genotypic value between homozygote and heterozygote,

$$\underset{\sim}{P} = \begin{pmatrix} \sigma_1^2 & 2q(1-q)a \\ 2q(1-q)a & 2q(1-q) \end{pmatrix}, \qquad \underset{\sim}{G} = \begin{pmatrix} h_1^2\sigma_1^2 & \\ & 2q(1-q)a \end{pmatrix}.$$

The genetic variance of the economically important trait contributed by this locus is thus $2q(1-q)a^2 = k\sigma_1^2$, say, where k is the proportion of the total variance contributed by the locus. The rate of response from individual selection using this locus would be

$$R = h_1^2\sigma_1 \left[1 + \frac{(1-h_1^2)^2 k}{2h_1^4}\right]$$

approximately, Neimann-Sörensen and Robertson (1961) and Smith (1967) giving similar approximations. With accurate knowledge of parameters, the proportional gain is thus $(1-h_1^2)^2 k/2h_1^4$.

Lack of knowledge of the true parameter values will lead to a loss of efficiency in the index. Using a Taylor's series expansion it can be shown that if the genotype has a rather small effect on the trait, the expected proportional loss, L, depends critically on the

estimation of the genotypic value, $\underline{a}$. In the limiting case where the genotype has no effect ($\underline{a} = 0$), it can be shown that the proportional loss is given approximately by

$$L = \frac{-q(1-q)(1-h_1^2)^2}{h_1^4\sigma_1^2} V(\hat{a}) \qquad (5)$$

using (2).

If $\underline{a}$ is calculated from the regression of x_1 on x_2 in a random sample of unrelated individuals, $V(\hat{a})$ is the variance of the regression coefficient and is

$$V(\hat{a}) = \sigma_1^2/2Tq(1-q) , \qquad (6)$$

where T is the total number of individuals measured and the residual variance is σ_1^2 since the true slope of the regression is zero. Hence from (5) and (6)

$$L = \frac{-(1-h_1^2)^2}{2Th_1^4} .$$

This, with the negative sign removed, is also the predicted proportional gain from the estimates of parameters. Thus, letting $u = (1-h_1^2)^2/2h_1^4$, it turns out that for small k the optimum proportional gain is ku, whereas the predicted benefit when $k = 0$ is u/T and the real loss is also u/T.

In practice, $\underline{a}$ may be estimated from the same set of data as are other parameters, and would be estimated as the pooled within-sire regression coefficient in a half-sib family structure. $V(\hat{a})$ will be higher than given by (6) since within progeny groups from two homozygous sires only two genotypes will be represented. The relevant

variances of x_1 and x_2 are within families; the variance of x_1 is reduced from σ^2 to $\sigma^2 - h_1^2\sigma_1^2/4$, and of x_2 from $2q(1-q)$ to $\frac{3}{2}q(1-q)$. If full sibs are used there is a further, corresponding reduction in both variances. With most sets of data (6) will be an underestimate, however, for sire family effects would be confounded with estimates of gene effects and the within-family analysis would be relatively more efficient.

DISCUSSION

Magnitude of gains and losses

In Part I (Sales and Hill, 1976) the incorporation of family information in an index to select for merit on a single trait was discussed. Except with very low heritabilities (or intra-class correlation of family members) the benefit from including the family mean after individual performance was less than 50%. Although predictions of absolute response, and to a lesser extent the relative magnitudes of response from index and individual information, were somewhat sensitive to errors in estimates of parameters, the response achieved (R*) was very insensitive to errors in the estimate of the only critical parameter, the intra-class correlation. It was found that if the intra-class correlation were estimated from data on 20 or more families of the size to be used subsequently in the selection programme, the proportional loss in response (L) would be less than 1% for most parameter values. In this paper the discussion has centred on additional information included in the index from a different trait. The benefits that might accrue from doing so depend, of course, on the parameters. In the examples of Table 1

these were up to about 40% of extra improvement; but Gjedrem (1967) gave examples where the gain was up to 300% for an important trait of low heritability ($h_1^2 = 0.1$). Prediction of responses are again found to be unreliable without very large initial samples of data, but, in contrast to the use of relatives' information, the gains achieved (R^*) can also be far from optimal. For example, with $h_1^2 = 0.2$, $h_2^2 = 0.5$ and uncorrelated traits, the proportional loss is roughly 60/T with families of size 16 (Table 1, Figure 2). Thus to reduce this loss to 1% requires that the number of families (s) satisfies $60/16s < 0.01$ or $s > 375$. Table 1 shows that somewhat fewer families would be required in the case where there was a real benefit, but many more families if their individual size was smaller. The greatest gains from using a secondary trait are obtained when the economically important trait has a low heritability, but in the latter case proportional losses are greater, being roughly of the order of one-quarter of the coefficient of variation of the heritability estimate of the important trait in the same experiment. The reason why the index is much more robust for family information appears to be that there the "secondary" measurement, the family mean, has a high positive correlation with breeding value, and providing it is giving positive weight is of benefit. By contrast, a secondary trait may be of no real benefit, or the signs used for the correlations may be incorrect.

Most emphasis has been given in this paper to the special case where the additional trait really contributes nothing useful, which is where wrong decisions are most easily made and likely to have most effect (see Figure 1). Although the obvious example of a model where the additional trait is of no value is where the second

trait is both genotypically and phenotypically uncorrelated with the economically important trait, the condition for the additional trait to be of no value is $r_G h_2 = r_P h_1$, i.e. the genetic and phenotypic regressions of trait 2 on trait 1 are equal. Thus a secondary trait which appears to have a useful correlation with the first could contribute nothing. It is therefore insufficient to base arguments about incorporation of a secondary trait solely on phenotypic information, which can easily be obtained; it is essential to have data on genotypes also and thus, of course, large scale preliminary experiments are required.

The analysis has dealt solely with one additional trait, but can readily be extended to more: the $\underset{\sim}{P}$ and $\underset{\sim}{G}$ matrices are increased to incorporate the additional traits. For any set of true parameters Monte Carlo simulation can be used, but it is difficult thereby to see a general pattern. By use of the Taylor's series approximation it can be shown that for the special case of mutually uncorrelated traits, each additional trait added to the index reduces the expected response by the same amount as if it were the first added; so, for identically distributed traits, the loss is proportional to the number added. Thus, if 10 independent biochemical variants are included in the index and none of them are associated with the trait of importance, the expected loss will be 10 times as great as if only one is included. When the traits or variants are correlated, this proportionality no longer holds.

Harris (1964) investigated in detail a model in which two traits had equal economic weights. It can be shown using the Taylor's approximation that when two traits are uncorrelated, the expected proportional loss is independent of the relative sizes of the economic

weights if the traits have the same heritability. Thus the proportional losses in Harris's and our models are, asymptotically, the same with traits of equal heritability. If, however, the traits are of unequal heritability, the proportional loss increases as the relative economic weight given to the trait of low heritability increases. Thus the case where the economically important trait is of low heritability and other traits are not of direct economic importance, the model to which we have given most attention, is that most sensitive to errors of estimation of parameters.

Partial indices

The standpoint of this paper has primarily been to consider the effect of adding a second trait when, with poor estimates of parameter values, an increase in response is invariably predicted, whether or not a real increase will actually be achieved. Often the reverse approach is used: when information has been collected on several traits with the intention of using all of them, analyses can be made of the effect of deleting traits from the index (e.g. Pig Industry Development Authority, 1965). These partial indices may be much cheaper to use since less data on individual animals are required, even if some response is sacrificed. Invariably deletion of any trait will be predicted to reduce the efficiency of the index, but if the estimates of parameters are poor, it is possible that the efficiency of the partial index might be higher than that of the full index.

Variance of predicted responses

The analysis has dealt only with average predicted benefits and losses, yet in planning, justifying and monitoring a breeding programme

some idea of the variance of the predicted response is required. This variance, $V(\hat{R})$ can be calculated from the usual approximation based on the first derivative terms in Taylor's series. Assuming $\hat{R}$ is a function of observables, y_i, for example mean squares and crossproducts in the analysis of variance,

$$V(\hat{R}) = \Sigma\Sigma(\partial\hat{R}/\partial y_i)(\partial\hat{R}/\partial y_j) \text{ cov } (y_i, y_j).$$

It turns out that $V(\hat{R})$ is a complicated function of the parameters, but when the true contribution of the second trait is small, $V(\hat{R})$ depends almost entirely on the parameters of the important trait. Then $V(\hat{R})$ is always rather higher than $V(\hat{h^2}\sigma) = V(\hat{R}_1)$, and its coefficient of variation exceeds that of the coefficient of variation of the heritability estimate, which is likely to be substantial unless the data are extensive.

Perhaps more important is the variance of the ratios of both the predicted and achieved responses from index selection on two traits versus selection only on the important trait. The ratios $\hat{R}/\hat{R}_1$ and R^*/R_1 are seen in Figure 1 to have very skewed distributions, especially when the second trait should contribute little or nothing when unity is a lower bound for $\hat{R}/\hat{R}_1$ and an upper bound for R^*/R_1^* with a concentration of points near unity; so even if a simple analytical formula for these variances had been obtained it would not have conveyed sufficient information. It has been found using Monte Carlo simulation that, especially for the zero correlation case, the deviations of the ratios from unity have roughly a chi-square distribution with one degree of freedom multiplied by a scalar constant $\pm\lambda$, where λ is defined in (2). For example, with $s = 100$, $n = 16$, $h_1^2 = 0.2$, $h_2^2 = 0.5$ and uncorrelated traits, λ is computed to be

0.037 by Taylor's series and in 400 replicates, $E(\hat{R}/\hat{R}_1) = 1.046$ and $E(R^*/R_1^*) = 0.961$. Using chi-square the predicted ratio would exceed 1.1 in 55 replicates, the observation was 46 replicates. With a doubling of the initial number of sire families, λ and the proportions change accordingly. The critical problem is that even in a design where mean losses are not large, there will be some samples of data which could lead to very large losses in efficiency if the index of two traits were used.

Removal of bias

There exist statistical procedures for removing bias from estimators where the bias is inversely proportional to the sample size. The most widely adopted is the "jackknife" technique (see Miller, 1974, for a recent review) in which sections of the data are analysed in turn. The jackknife procedure was tried on simulated half-sib family data for two uncorrelated traits, such as used to produce Figure 1a (F.H.L. Gilchrist and W.G. Hill, unpublished). Each sire family was omitted in turn and the data combined in the standard manner. The resultant estimator was usually somewhat biassed in the opposite direction (i.e. $E(\hat{R}/\hat{R}_1) < 1$), but the main problem was that it very markedly increased the sampling variance of $\hat{R}/\hat{R}_1$. In some situations problems were also encountered if omission of one family led to negative heritability estimates in the sub-sample and there was no obvious way of coping with this. The attempts to use the jackknife were abandoned. Perhaps the main value of this study was to draw more attention to the problem of variation in the predicted responses and particularly the extreme values discussed above.

Prior information

Largely for analytical simplicity it has been assumed that the only information available on the traits (and particularly the secondary traits) is contained in the data available for analysis. In most practical situations there is likely to be prior information, or at least prior prejudices perhaps based on physiological arguments, about the relevant correlations and heritabilities. How should this be taken into account? If the data are of equivalent type they can, of course, be incorporated with the new data, or in the unlikely event that the prior information were sufficiently well defined that a prior probability distribution could be constructed, Bayesian methods could be used to incorporate it.

An alternative viewpoint is to give more weight to the prior economic information when there is doubt about the accuracy of the estimates of genetic parameters. The extreme approach is to use a so-called "base index" in which the economic vector (expressed in phenotypic standard deviations) is used instead of the computed index, a procedure discussed by Mao (1971) in a similar situation to that discussed here. In our example, the base index is simply selection on the important trait alone. A sophistication of the procedure is to regress the computed index to the base index, but this methodology has not been worked out.

Need for decision rules

Ideally, this paper would conclude with a rule on how to make decisions about whether or not to include information on additional traits in the index, or to exclude information to construct a partial index. This problem is analogous to that of including extra variables in a multiple regression equation and one would hope to obtain some guidance from that field. The usual operational procedure in multiple

regression is to include an extra variable when it contributes significantly more than expected from the error variance, and the increase in accuracy (measured as the multiple correlation coefficient, coefficient of determination or residual standard deviation) is sufficient to justify the cost of taking the additional measurements in the future. But such a procedure implies that, with some probability, measurements will be included which are given a high weight in the regression, yet contribute little or nothing and markedly reduce the real as opposed to estimated correlation coefficient. This procedure therefore has its deficiencies, even if it could readily be extended to the case of selection indices where significance tests are much less straightforward.

Clearly further work is required on decision rules. It is hoped that this paper has illustrated that there is a selection index problem worth solving.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the Meat and Livestock Commission. We are grateful to Marjorie McEwan for computational assistance, and to Alan Robertson, David Sales and Charles Smith for helpful comments.

REFERENCES

GJEDREM, T. 1967. Selection indices compared with single trait selection. I. The efficiency of including correlated traits. Acta Agric. Scand. 17: 263-268.

HARRIS, D.L. 1964. Expected and predicted progress from index selection involving estimates of population parameters. Biometrics 20: 46-72.

HARTLEY, H.O. and HARRIS, D.L. 1963. Monte Carlo computations in normal correlation problems. J. Assoc. Comp. Mach. 10: 302-306.

MAO, I.L. 1971. The effect of parameter estimation errors on the efficiency of index selection and on the accuracy of genetic gain prediction. Ph.D. Thesis, Cornell Univ., Ithaca, New York.

MILLER, R.G. 1974. The jackknife - a review. Biometrika 61: 1-15.

NEIMANN-SØRENSEN, A. and ROBERTSON, A. 1961. The association between blood groups and several production characters in three Danish cattle breeds. Acta Agric. Scand. 11: 163-196.

PIG INDUSTRY DEVELOPMENT AUTHORITY. 1965. Combined testing. Recommendations by the statistics section for the selection index. Mimeograph DA188.

ROBERTSON, A. 1959. Experimental design in the estimation of genetic parameters. Biometrics 15: 219-226.

SALES, J. and HILL, W.G. 1976. Effect of sampling errors on efficiency of selection indices. I. Use of information from relatives for single trait improvement. Anim. Prod. 22: 1-17.

SMITH, C. 1967. Improvement of metric traits through specific genetic loci. Anim. Prod. 9: 349-358.

23

Linkage disequilibrium in finite populations

by

William G. Hill and Alan Robertson

Theoretical and Applied Genetics 38, 226–231 (1968)

Linkage Disequilibrium in Finite Populations

W. G. HILL and ALAN ROBERTSON*

Institute of Animal Genetics, Edinburgh, 9.

Summary. A theoretical investigation has been made of the influence of population size (N) and recombination fraction (c) on linkage disequilibrium (D) between a pair of loci. Two situations were studied: (i) where both loci had no effect on fitness and (ii) where they showed heterozygote superiority, but no epistacy.

If the populations are initially in linkage equilibrium, then the mean value of D remains zero with inbreeding, but the mean of D^2 increases to a maximum value and decreases until fixation is reached at both loci. The tighter the linkage and the greater the selection, then the later is the maximum in the mean of D^2 reached, and the larger its value. The correlation of gene frequencies, r, in the population of gametes within segregating lines was also studied. It was found that, for a range of selection intensities and initial gene frequencies, the mean value of r^2 was determined almost entirely by $N\,c$ and time, measured proportional to N.

The implication of these results on observations of linkage disequilibrium in natural populations is discussed.

Most of the mathematical theory of linkage has been developed for populations which are sufficiently large that a deterministic model can be used. In these large populations, which are not undergoing selection, the theory of the rate of approach to linkage equilibrium is well worked out, and it is known that populations in equilibrium remain in that state (GEIRINGER, 1944; BENNETT, 1954). More recent work has been devoted to the effect of selection on linkage disequilibrium in very large populations. LEWONTIN and KOJIMA (1960) showed that epistacy was necessary for linkage disequilibrium to be maintained in a selected population in which gene frequencies are at equilibrium. However in a population in which there are directional changes of gene frequency resulting from artificial selection, some linkage disequilibrium may be observed if there is no epistacy of gene action on the selected character (NEI, 1963; FELSENSTEIN, 1965), but not if the selective values at each loci combine in a multiplicative manner (FELSENSTEIN, 1965).

For finite unselected random mating populations, expressions for changes in linkage disequilibrium have been given by KIMURA (1963) and by HILL and ROBERTSON (1966). WRIGHT (1933) had previously derived formulae for the proportion of recombinants at final fixation. In this paper we shall mainly consider the fate of a pair of linked loci, which we may observe in a number of replicate lines drawn from a large population initially in linkage equilibrium. If the loci have no effect on fitness, then over the average of all replicates these loci will remain in equilibrium, but as a result of genetic sampling the disequilibrium will not be zero in each line. In other words, the variance of the linkage disequilibrium, D, will not be zero, though the mean will be. We shall evaluate this variance, and show that it can be of an order of magnitude similar to that of the variance of gene frequencies after some generations of inbreeding. We shall study the case of neutral genes in greatest detail, and then extend the results to include heterozygote advantage at each locus, but with no epistacy. The results may therefore have some bearing on the estimation and interpretation of linkage disequilibrium in natural populations.

* Member of the A.R.C. Unit of Animal Genetics.

Disequilibrium Between Neutral Loci

We consider two loci with alternative alleles A_1, A_2 and B_1, B_2 which have no effect on fitness, and we let p and q be the frequencies of A_1 and B_1 respectively. Linkage disequilibrium is commonly measured by the determinant D, given by

$$D = f(A_1\,B_1)\,f(A_2\,B_2) - f(A_1\,B_2)\,f(A_2\,B_1)$$

where f denotes the appropriate gametic frequency.

Using E to denote expectation, the recurrence equation for the mean of D after t generations of random mating with no selection is

$$E(D_t) = (1 - c)\,(1 - 1/2\,N)\,E(D_{t-1})\,, \qquad (1)$$

where N is the effective population size and c the cross-over distance (HILL and ROBERTSON, 1966). If c and $1/2\,N$ are sufficiently small that their product can be ignored

$$\begin{aligned} E(D_t) &= (1 - c - 1/2\,N)\,E(D_{t-1}) \\ &= D_0\,e^{-(2\,N\,c+1)\,t/2\,N}\,, \quad \text{approximately} \qquad (2) \end{aligned}$$

if the population size is constant. In general, if N is large and c is of order $1/N$ or less, changes in the distribution of gametic frequencies can be approximated in a continuous model using a diffusion equation. Under these assumptions it can be shown that the pattern of change in gametic frequencies is a function of only the initial conditions p_0, q_0 and D_0 and of the product $N\,c$, if time is expressed on a scale proportional to N (HILL and ROBERTSON, 1966). Equation (2) is clearly of this form.

Changes in the average value of D^2 can be obtained using a moment generating matrix (ROBERTSON, 1952). Let y be a column vector of moments with dimension three, and elements

$$\begin{aligned} y_1 &= E\,[p\,(1 - p)\,q\,(1 - q)]\,, \\ y_2 &= E\,[D\,(1 - 2\,p)\,(1 - 2\,q)]\,, \\ y_3 &= E[D^2]\,. \end{aligned}$$

If there is no crossing over, changes in these moments in successive generations can be obtained by taking expectations over the multinomial distribution of gametic frequencies with index n, where $n = 2\,N$, and rearranging the results in terms of y_1, y_2, and y_3. Denoting by $y_{(t)} = (y_{i(t)})$ the vector of moments at

some generation t, it can be shown that

$$y_{1(t+1)} = \left(1-\frac{1}{n}\right)^2 y_{1(t)} + \frac{1}{n}\left(1-\frac{1}{n}\right)^2 y_{2(t)} + \frac{2}{n^2}\left(1-\frac{1}{n}\right) y_{3(t)}$$

$$y_{2(t+1)} = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)^2 y_{2(t)} + \frac{4}{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) y_{3(t)}$$

$$y_{3(t+1)} = \frac{1}{n}\left(1-\frac{1}{n}\right) y_{1(t)} + \frac{1}{n}\left(1-\frac{1}{n}\right)^2 y_{2(t)} + \left(1-\frac{1}{n}\right)\left[\frac{1}{n^2}+\left(1-\frac{1}{n}\right)^2\right] y_{3(t)} .$$

When crossing over occurs, the average values of gene frequencies are unchanged, terms in D are multiplied by a factor $(1-c)$ (c. f., equation (1)), and terms in D^2 by a factor $(1-c)^2$. Thus with crossing over followed by sampling, we obtain the following transition relationship:

$$y_{(t+1)} = \begin{pmatrix} \left(1-\frac{1}{n}\right)^2 & \frac{1}{n}\left(1-\frac{1}{n}\right)^2(1-c) & \frac{2}{n^2}\left(1-\frac{1}{n}\right)(1-c)^2 \\ 0 & \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)^2(1-c) & \frac{4}{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)(1-c)^2 \\ \frac{1}{n}\left(1-\frac{1}{n}\right) & \frac{1}{n}\left(1-\frac{1}{n}\right)^2(1-c) & \left(1-\frac{1}{n}\right)\left[\frac{1}{n^2}+\left(1-\frac{1}{n}\right)^2\right](1-c)^2 \end{pmatrix} y_{(t)} \qquad (3)$$

$$= \boldsymbol{M}\, y_{(t)} ,$$

where $\boldsymbol{M}$ is termed the moment generating matrix and is independent of t, hence $y_{(t)} = \boldsymbol{M}^t y_{(0)}$.

We shall discuss in detail the case of initial linkage equilibrium, where $y'_{(0)} = (p_0(1-p_0)\, q_0(1-q_0)\, 0\, 0)$. Then $E(D^2)/[p_0(1-p_0)\, q_0(1-q_0)]$ is independent of the initial frequencies at the two loci, and under the continuous model assumptions will be a function only of $N\,c$ and time expressed proportional to N. Since $E(D) = 0$ if $D_0 = 0$, $E(D^2)$ measures the variance of D.

With complete linkage ($c = 0$) and large N an explicit solution for $E(D^2)$ can be obtained. This involves diagonalization of the moment generating matrix, and the derivation and general solution are given in the appendix. With initial equilibrium, the solution is

$$E(D^2) = \frac{1}{15} p_0(1-p_0)\, q_0(1-q_0)\,[6(1-F) - 5(1-F)^3 - (1-F)^6] ,$$

where F is the inbreeding coefficient.

Some results for $E(D^2)/[p_0(1-p_0)\, q_0(1-q_0)]$ for initial equilibrium are plotted in Figure 1, with time measured as $F = 1 - e^{-t/2N}$. The graphs were computed by repeated iteration of the matrix, $\boldsymbol{M}$, on to the vector $y_{(t)}$ using a population size of $N = 16$. When $c = 0$, $E(D^2)$ reaches a maximum of 0.165 $p_0(1-p_0)\, q_0(1-q_0)$ when $F = 0.4$, or $t = N$ generations approximately. With recombination, the maximum value of $E(D^2)$ is lower and is attained earlier. For example, for $N\,c = 1/4$, $E(D^2)$ reaches 0.14 $p_0(1-p_0)\, q_0(1-q_0)$ when $F = 0.31$ or $t = 0.75\,N$ generations.

We have made use of the generalisation derived from the continuous model that c only enters into the results in the form of $N\,c$ and that the time scale in generations is proportional to N. This was checked in the calculations and in Table 1 we present some of the results referring to the maximum values of D^2 reached. The table gives the observed maximum and the time in generations when it occurred.

Except for the smallest values of N and $N\,c = 4$ there appears to be sufficiently good agreement between the results obtained with different values of N for us to use this generalisation.

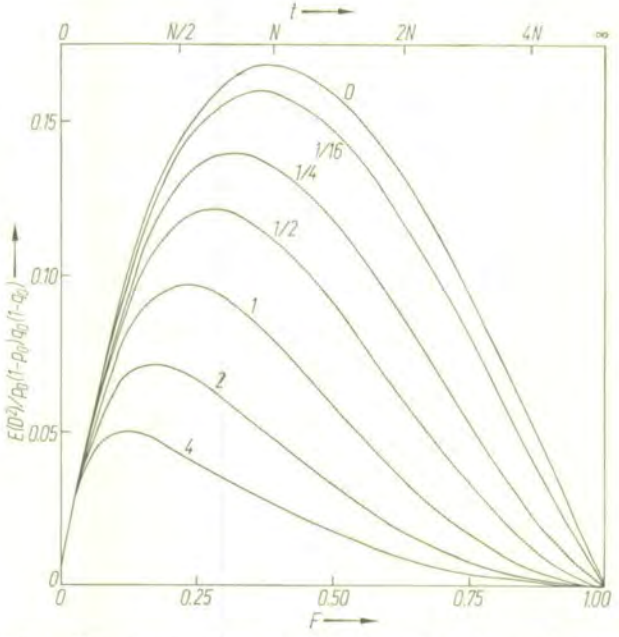

Fig. 1. The mean value of $D^2/[p_0(1-p_0)\, q_0](1-q_0)$ over segregating and non segregating lines for several values of $N\,c$ and no selection

Table 1. *The maximum value of $E(D^2)$ and the time in generations to reach it for different combinations of N and $N\,c$*

	N					$N\,c$
	8	16	32	64	$\to \infty$	
$D^2_{max}/[p_0(1-p_0)\, q_0(1-q_0)]$	.1708	.1678	.1663	.1656	.1649	0
t	8	16	32	64	1.0016 N	
$D^2_{max}/[p_0(1-p_0)\, q_0(1-q_0)]$	.1054	.0969	.0931	.0913	—	1
t	4	8	17	34	—	
$D^2_{max}/[p_0(1-p_0)\, q_0(1-q_0)]$	.0636	.0503	.0451	.0428	—	4
t	2	4	8	17	—	

The product of the variances in gene frequencies at the two loci, $p\,(1-p)\,q\,(1-q)$, is also affected by the degree of linkage. Starting with equilibrium, we have for $c = 0$ from the appendix that

$$E\,[p\,(1-p)\,q\,(1-q)\,] = \frac{1}{15}p_0\,(1-p_0)\,q_0\,(1-q_0) \times [6\,(1-F) + 10\,(1-F)^3 - (1-F)^6]\,.$$

With independent loci, the variance at each locus is proportional to $1 - F$, and their product is proportional to $(1 - F)^2$. However, with complete linkage, the product of the variances is

$$\frac{1}{15}\left[\frac{6}{1-F} + 10\,(1-F) - (1-F)^4\right]$$

times that with independence for all starting frequencies. This ratio is 1 at $F = 0$, rises to 1.13 at $F = 0.5$, 4.07 at $F = 0.9$ and becomes infinitely large as F approaches one.

Disequilibrium in Populations Segregating at Both Loci

We have developed the analysis so far in terms of the average values of D^2 computed over all replicate lines. But in many replicates one or other locus will become fixed after a few generations and in these D is zero. If we observe linkage disequilibrium between a pair of loci in natural populations, it can only be among those still segregating at both loci. We therefore need to describe the behaviour of the linkage disequilibrium within such lines. When $c = 0$, the average value of D^2 within lines still segregating at both loci, denoted D_s^2, can be obtained by dividing $E(D^2)$ from equation (3) by the proportion of lines still segregating. The latter can be calculated by series summation from formulae by KIMURA (1955), regarding the four gamete types as four alleles at a single locus. In the limiting case with $c = 0$, as F approaches 1 only lines in which two gametes segregate will remain. In those still segregating for both loci, gametes must either be entirely in the repulsion or entirely in the coupling phase. Therefore if we assume a final uniform continuous distribution of gametic frequencies ($A_1\,B_2$, $A_2\,B_1$) or ($A_1\,B_1$, $A_2\,B_2$) as will be true if N is large, it can be shown that D_s^2 approaches 1/30 for complete linkage as the inbreeding coefficient approaches one. This final value is independent of the initial conditions p_0, q_0 and D_0.

The values of both D^2 and D_s^2 depend during inbreeding on the initial frequencies, and we have found that a more useful statistic for lines segregating at both loci is the square of the correlation, r, of gene frequencies in the population of gametes, where $r = D/[p\,(1-p)\,q\,(1-q)]^{1/2}$. The expectation of r or r^2 is computed by averaging only over such lines. When there is initial equilibrium $E(r) = 0$. Changes in $E(r^2)$ with level of inbreeding were obtained by Monte Carlo simulation, using the same procedure as in our earlier work, but excluding selection (HILL and ROBERTSON, 1966). A population size of $N = 8$ was used and 10000 replicates were run for each level of recombination. Results are shown in Figure 2 for $p_0 = q_0 = 0.5$ and $D_0 = 0$. As replicate lines become fixed in the later generations, the sampling variances of the estimates of $E(r^2)$ increase, so results are plotted for only 48 generations ($F = 0.95$).

When there is complete linkage ($N\,c = 0$), $E(r^2)$ eventually reaches unity as all lines approach fixation. It is interesting that $E(r^2)$ and F are approximately equal to each other when $N\,c = 0$. When there is recombination, $E(r^2)$ approaches a limiting value dependent on $N\,c$ as F approaches one, the limit being reached earlier and at a lower level, the less tight the linkage. It is difficult to estimate the limit of $E(r^2)$ accurately when $N\,c$ is small because few lines are still segregating when $E(r^2)$ has reached a stable value.

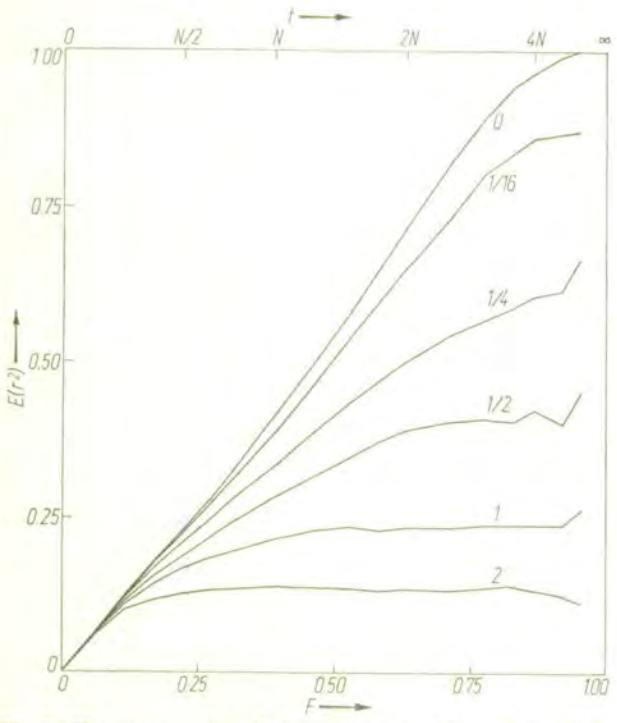

Fig. 2. The mean value of r^2 among segregating lines, $E(r^2)$, for several values of $N\,c$ with no selection

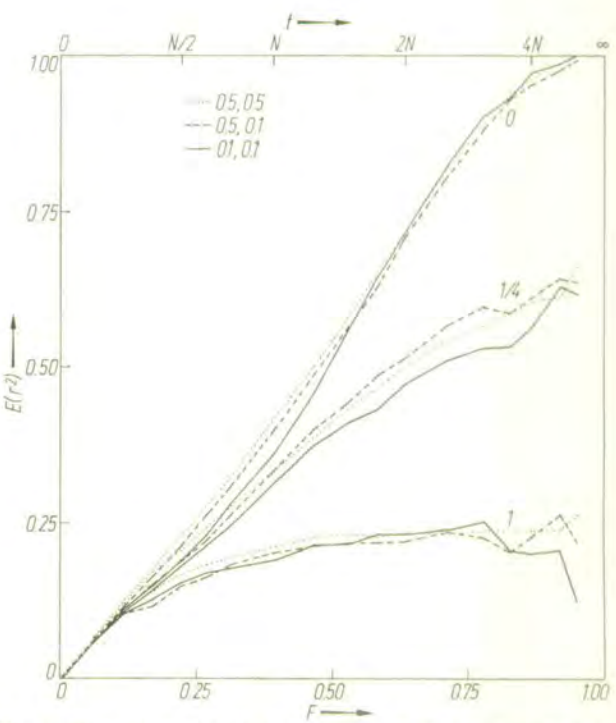

Fig. 3. The effect of initial frequency on $E(r^2)$ with no selection

The influence of initial frequency on $E(r^2)$ when there is no selection is shown in Figure 3. Three sets of initial frequencies are compared: (i) $p_0 = q_0 = 0.5$, (ii) $p_0 = 0.1$, $q_0 = 0.5$ and (iii) $p_0 = q_0 = 0.1$, each with $D_0 = 0$. It appears from Figure 3 that $E(r^2)$ is not very sensitive to changes in the initial frequency, and is mostly determined by $N\,c$ and time (measured as a function of N). As the inbreeding coefficient approaches unity, $E(r^2)$ depends only on the steady state distribution of gamete frequencies within the segregating lines, and is independent of the initial conditions.

Disequilibrium Between Loci Having Heterozygote Superiority

Our discussion has been restricted so far to the situation where there is no selection maintaining segregation. But genetic variation will be maintained for longer periods of time in small populations at loci in which the heterozygote has superior fitness to either homozygote, unless the homozygotes differ widely from each other in fitness (ROBERTSON, 1962; ROBERTSON and HILL, 1968). If there is no epistacy between these loci, selection will not cause linkage disequilibrium directly. But we must expect to find some disequilibrium in small populations as a result of genetic sampling. We can therefore predict that for pairs of loci each having heterozygote advantage, but not interacting with each other, we will have $E(D) = 0$, but $E(D^2) \neq 0$, just as for neutral genes.[1]

Let us assume that the relative selective advantages are as follows:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$1 - r_1 - s_1$	$1 - r_1$	$1 - r_1 - s_2$
A_1A_2	$1 - s_1$	1	$1 - s_2$
A_2A_2	$1 - r_2 - s_1$	$1 - r_2$	$1 - r_2 - s_2$

The equilibrium gene frequencies for large populations are given by $\bar{p} = r_2/(r_1 + r_2)$ at the A locus, and $\bar{q} = s_2/(s_1 + s_2)$ at the B locus. The change in gene frequency at the A locus in one generation is given by $\delta p = -(r_1 + r_2)\, p\,(1 - p)\,(p - \bar{p})$, with a similar equation for locus B, where squared terms in selective values are ignored. On a continuous model it can be shown that, on a time scale proportional to N, the inbreeding and selection process is a function of only $\bar{p}$, $\bar{q}$, $N\,c$, $N\,(r_1 + r_2)$ and $N\,(s_1 + s_2)$ for a given set of initial conditions p_0, q_0 and D_0. No explicit solutions for this model could be obtained, so our Monte Carlo programme was modified to include selection for heterozygotes. The number of replicates used for each set of parameters depended on the rate of fixation observed, and was chosen so that roughly the same number of replicates were segregating at both loci after $4\,N$ generations as for the case of no selection with $p_0 = q_0 = 0.5$, and 10000 replicates. All simulation was done with $N = 8$, except for one example with $N\,(r_1 + r_2) = N\,(s_1 + s_2) = 4$ and $\bar{p} = \bar{q} = 0.5$ which was also run with $N = 16$ (Figure 4).

Selection is most effective in maintaining heterozygosity when the equilibrium gene frequency is one-half (ROBERTSON 1962; ROBERTSON and HILL, 1968).

[1] $E(D) = 0$ with heterozygote superiority at both loci only if $\bar{p} = 0.5$ and/or $\bar{q} = 0.5$.

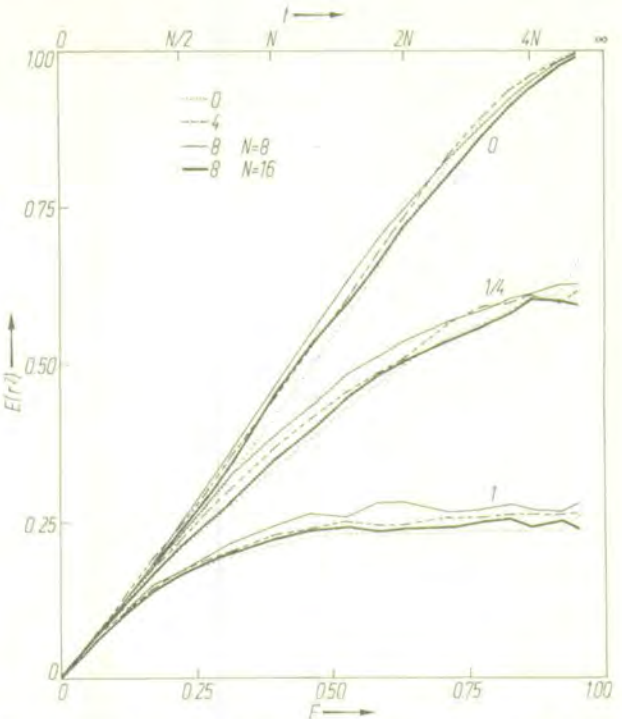

Fig. 4. The effect of selection for heterozygotes on $E(r^2)$ for $N\,(r_1 + r_2) = N\,(s_1 + s_2) = 0$, 4 and 8 and $p_0 = q_0 = \bar{p} = \bar{q} = 0.5$. Populations were simulated with both $N = 8$ and $N = 16$ for $N\,(r_1 + r_2) = 8$, otherwise $N = 8$

We shall therefore discuss this situation ($\bar{p} = \bar{q} = 0.5$) in most detail, and for simplicity assume that $N\,(r_1 + r_2) = N\,(s_1 + s_2)$.

Such selection has two related consequences which are relevant here. Firstly, the rate of fixation may be greatly reduced and secondly, the gene frequency distribution amongst unfixed lines becomes more concentrated around the equilibrium frequencies as selection becomes more intense. We found that $E(D^2)$ over all populations was increased by selection for heterozygotes. This appears to be mainly due to retardation of fixation as the effect on $E(D_s^2)$ in segregating lines, though present, is small.

However, on examining $E(r^2)$, we found that this reaches a limiting value which is little influenced by the intensity of selection (Figure 4). Further it appears that about the same level of $E(r^2)$ is reached when the equilibrium frequency in large populations is not 0.5 (Figures 5 and 6). The curves of $E(r^2)$ against F or t/N, are then dependent almost only on $N\,c$. The limiting value of $E(r^2)$ appears to approach $1/4\,N\,c$ as $N\,c$ increases. A crude derivation can be obtained by equating the loss in $E(D^2)$ each generation ($2\,c\,E(D^2)$ approximately) with the gain due to sampling ($p\,(1 - p)\,q\,(1 - q)/2\,N$). The second term in the vector y is then small because gene frequencies are close to 0.5.

Discussion

Many workers are now investigating polymorphisms in natural populations, and opportunities will no doubt arise for measuring the linkage disequilibrium between the segregating loci observed. We have used the square of the correlation of gene frequencies, r^2, as our statistic, which has a known sampling distribution when the true value is zero.

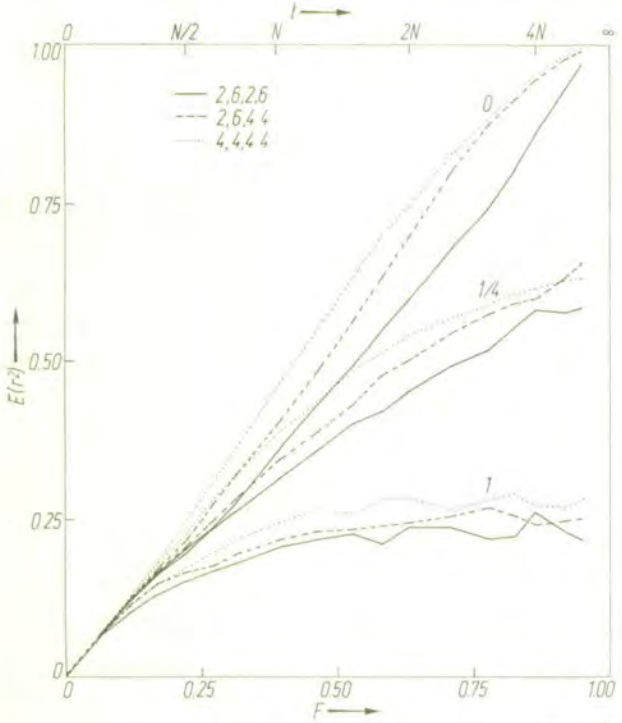

Fig. 5. The effect of equilibrium frequency on $E(r^2)$ with $N(r_1 + r_2) = N(s_1 + s_2) = 4$ and $p_0 = \overline{p}$, $q_0 = \overline{q}$

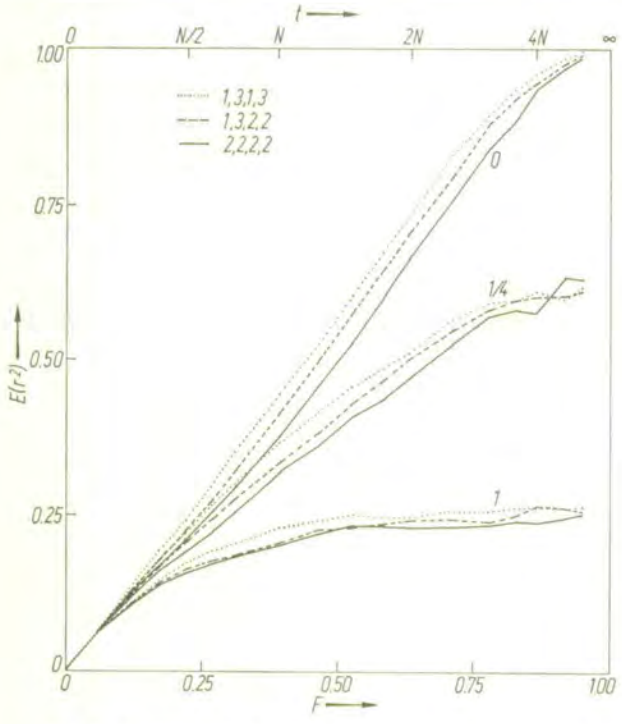

Fig. 6. As Figure 5, but $N(r_1 + r_2) = N(s_1 + s_2) = 8$

If a sample of T individuals are taken from the population, $T r^2$ is then distributed as χ^2 with one degree of freedom.

However our results show that when a significant departure from equilibrium is observed in a small population, we must be cautious about concluding that this is due to natural selection. Several models with interaction of selective advantage between the loci have been investigated in infinite populations. For example, Lewontin (1964) studied a two locus model with heterozygote advantage and epistacy, which had relative fitnesses as follows:

	A_1A_1	A_1A_2	A_2A_2
B_1B_1	.4	.6	.3
B_1B_2	.6	1.0	.5
B_2B_2	.5	.7	.4

From his Table 4 we can compute the values of r^2 reached at equilibrium. Two stable situations were possible, in which there was a final excess of either coupling or repulsion phases; we shall use only the latter. The results for the model were:

c	0	.01	.02	.04	.08
r^2	1.000	.799	.601	.221	.002

With no selection, or selection for heterozygotes with no epistacy, the mean value of r^2 within the segregating populations would reach the following approximate values, assuming a population size of $N = 25$ was maintained for many generations:

c	0	.01	.02	.04	.08
$E(r^2)$	1.00	.62	.41	.25	.12

These results correspond to $N c$ values of 0, .25, .5, 1 and 2. Thus two completely different processes lead to superficially similar results. It can be argued that $N = 25$ is much too small to represent a natural population. However, Lewontin's selective advantages with differences of factors of two at a single locus may be considered unrealistically large.

The model we have used may also be criticised because of the assumption of constant population size. However this does not effect the qualitative aspects of our results. Any restriction of population size may cause disequilibrium as a result of genetic sampling, and the return to equilibrium will be slow if the loci are tightly linked.

Zusammenfassung

Es wurde eine theoretische Untersuchung über den Einfluß der Populationsgröße (N) und der Rekombinationsfraktion (c) auf das Koppelungs-Ungleichgewicht (D) zwischen einem Paar von Loci angestellt. Die nachfolgenden zwei Situationen wurden studiert:

1. Beide Loci haben keinen Effekt auf die Fitness.

2. Die Heterozygoten zeigen Überlegenheit, jedoch keine Epistasie.

Befinden sich die Populationen in einem ursprünglichen Koppelungsgleichgewicht, so bleibt der mittlere Wert von D bei Inzucht gleich null, jedoch steigt das Mittel von D^2 bis zu einem Maximalwert und fällt, bis die Fixierung an beiden Loci erreicht worden ist. Je enger die Koppelung und je stärker die Selektion ist, desto später wird das Maximum im Mittel von D^2 erreicht und desto größer ist sein Wert. Ferner wurde die Korrelation von Genfrequenzen, r, in der Population von Gameten innerhalb spaltender Linien untersucht. Es wurde gefunden, daß der mittlere Wert von r^2 für einen Bereich von Selek-

tionsintensitäten und ursprünglichen Genfrequenzen praktisch vollkommen bestimmt wurde durch Nc und die Zeit, gemessen proportional zu N. Abschließend wird die Bedeutung dieser Ergebnisse für Beobachtungen von Koppelungs-Ungleichgewichten in natürlichen Populationen diskutiert.

References

1. Bennett, H. J.: On the theory of random mating. Ann. Eugenics **18**, 311–317 (1954). – 2. Felsenstein, J.: The effect of linkage on directional selection. Genetics **52**, 349–363 (1965). – 3. Geiringer, H.: On the probability theory of linkage in Mendelian heredity. Ann. Math. Statist. **15**, 25–57 (1944). – 4. Hill, W. G., and A. Robertson: The effect of linkage on limits to artificial selection. Genet. Res. **8**, 269–294 (1966). – 5. Kimura, M.: Random drift in a multi-allelic locus. Evolution **9**, 419–435 (1955). – 6. Kimura, M.: A probability method for treating inbreeding systems, especially with linked genes. Biometrics **19**, 1–17 (1963). – 7. Lewontin, R. C.: The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49**, 49–67 (1964). – 8. Lewontin, R. C., and K. Kojima: The evolutionary dynamics of complex polymorphisms. Evolution **14**, 458–472 (1960). – 9. Nei, M.: Effect of selection on the components of genetic variance. In: Statistical Genetics and Plant Breeding, ed. by W. D. Hanson and H. F. Robinson. Publ. 982, National Academy of Sciences, National Research Council, Washington, D.C., pp. 501–515 (1963). – 10. Robertson, A.: The effect of inbreeding on the variation due to recessive genes. Genetics **37**, 189–207 (1952). – 11. Robertson, A.: Selection for heterozygotes in small populations. Genetics **47**, 1291–1300 (1962). – 12. Robertson, A., and W. G. Hill: The effects of inbreeding at loci with heterozygote advantage. Genetics (submitted for publication 1968). – 13. Wright, S.: Inbreeding and recombination. Proc. Nat. Acad. Sci., Wash. **19**, 420–433 (1933).

Appendix: Diagonalisation of the Moment Generating Matrix with Complete Linkage

We use well known theory to find for the matrix M the scalar latent roots λ_1, λ_2 and λ_3 and their associated latent vectors v_1, v_2, and v_3 of dimension 3 such that

$$M v_i = v_i \lambda_i , \qquad i = 1, 2, 3 .$$

Thus, if we let $\mathbf{\Lambda}$ be a 3×3 diagonal matrix of the latent roots λ_i and let $V = (v_1\, v_2\, v_3)$ be the 3×3 matrix of latent vectors, we have

$$M V = V \mathbf{\Lambda}$$

and

$$M = V \mathbf{\Lambda} V^{-1}$$

also

$$M^2 = V \mathbf{\Lambda} V^{-1} \cdot V \mathbf{\Lambda} V^{-1} = V \mathbf{\Lambda}^2 V^{-1}$$

and so on.

To obtain the moments $y_{(t)}$ we require

$$y_{(t)} = M^t y_{(0)}$$

which is given by

$$y_{(t)} = V \mathbf{\Lambda}^t V^{-1} y_{(0)} \qquad (1A)$$

and needs only scalar multiplication to evaluate $\mathbf{\Lambda}^t$.

With complete linkage, M is given by setting $c = 0$ in equation (3) of the text and its latent roots and vectors are easily obtained. These are

$$\lambda_1 = 1 - \frac{1}{n}, \qquad \lambda_2 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right),$$

$$\lambda_3 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right)$$

and

$$V = \begin{pmatrix} 1 & -2 & 1 \\ \frac{n-2}{n-1} & 2 & -4 \\ 1 & 1 & 1 \end{pmatrix}.$$

Since the inbreeding coefficient, F, equals $1 - \left(1 - \frac{1}{n}\right)^t$, it follows that for large n

$$\lambda_1^t = 1 - F , \quad \lambda_2^t = (1 - F)^3 , \quad \lambda_3^t = (1 - F)^6 ,$$

approximately,

$$\text{and } V = \begin{pmatrix} 1 & -2 & 1 \\ 1 & 2 & -4 \\ 1 & 1 & 1 \end{pmatrix}, \text{ approximately.}$$

The inverse of V is then

$$V^{-1} = \frac{1}{15}\begin{pmatrix} 6 & 3 & 6 \\ -5 & 0 & 5 \\ -1 & -3 & 4 \end{pmatrix}$$

If there is initial equilibrium (a restriction not required by the preceding theory)

$$y'_{(0)} = p_0 (1 - p_0)\, q_0 (1 - q_0)\, (1\ 0\ 0)$$

and substitution into (1A) gives the result

$$y_{(t)} = \frac{1}{15} p_0 (1 - p_0)\, q_0 (1 - q_0) \times$$

$$\times \begin{pmatrix} 6(1-F) + 10(1-F)^3 - (1-F)^6 \\ 6(1-F) - 10(1-F)^3 + 4(1-F)^6 \\ 6(1-F) - 5(1-F)^3 - (1-F)^6 \end{pmatrix}.$$

V/12/6 0,210
Printed in Germany

24

Maintenance of segregation at linked genes in finite populations

by

William G. Hill

JAPAN. J. GENETICS Vol. 44, Suppl. 1: 144-151 (1969)

MAINTENANCE OF SEGREGATION AT LINKED GENES IN FINITE POPULATIONS[1)]

W. G. HILL

Institute of Animal Genetics, Edinburgh 9, United Kingdom.

There have been several theoretical studies of the joint effects of linkage and selection in evolution in which population size has been assumed to be infinitely large so that deterministic models could be used. Inbreeding has been included, but in a large population by means of random selfing and outcrossing (Jain and Allard 1966). Algebraic difficulties have necessitated restriction to two loci, although Lewontin (1964) has simulated models with more. The theory has been reviewed by Bodmer and Parsons (1962) and more recently by Bodmer and Felsenstein (1967). An aspect which has received particular attention is the influence of linkage on gene frequencies and gametic frequencies at equilibrium. Lewontin and Kojima (1960) and Bodmer and Felsenstein (1967) have shown that, if there is no epistasis, linkage does not affect the final equilibrium of the population, which has a unique stable position with D (the linkage disequilibrium determinant) = 0. When epistasis is present linkage has to be fairly tight for there to be any effect on final equilibrium, when stable equilibria of gametic frequencies with $D \neq 0$ may be found. With a multiplicative model of fitness there is an equilibrium at $D = 0$, but this is unstable if there is very tight linkage (Bodmer and Felsenstein 1967).

Studies of the effects of linkage on selection response in closed finite populations have mostly been concerned with directional selection and have used Monte Carlo methods. Using two locus models Latter (1966) and Hill and Robertson (1966) have derived theoretical approximations to explain some of the effects of linkage with directional selection. More recently Karlin and McGregor (1968) and Hill and Robertson (1968) have drawn attention to the importance of genetic drift in causing linkage disequilibrium in finite populations, even between neutral loci. Hill and Robertson (1968) and Ohta and Kimura (1969) have derived exact formulae for the expected value of D^2 between neutral loci, where D has an expected value of zero if there is initial equilibrium. Hill and Robertson (1968) also used Monte Carlo methods to estimate r^2, the square of the correlation of gene frequencies between pairs of linked heterotic loci. In terms of this parameter, r^2, similar amounts of disequilibrium were found between pairs of neutral and between pairs of heterotic loci at which the homozygotes have similar fitness.

Sved (1968), Levin (1968) and the present author (Hill 1968) have noted that fixation of heterotic loci by drift is retarded if they are tightly linked. Sved gave some theoretical arguments and used Monte Carlo simulation with a model of 180 identical loci with symmetric fitness (*i.e.* homozygotes having the same fitness) which were equally spaced on a single chromosome. In this paper a model of only two loci will be investigated in more detail. Essentially the same phenomena of retarded fixation are found with two loci as with many loci, yet the simpler model requires less computer time for

1) Modified from talk entitled "Population dynamics of linked genes in finite populations" presented in a Small Symposium at XII International Congress of Genetics, Tokyo, August, 1968.

any choice of parameters. An attempt is also made to evaluate the importance of epistasis between linked heterotic loci in small populations by comparing the rates of fixation for several models with epistasis with one in which the fitnesses are additive between loci, since these have very different properties in infinitely large populations.

MODEL

The necessary parameters may be summarised as follows

Loci	A	B
Alleles	A_1, A_2	B_1, B_2
Frequency, f	$f(A_1) = p$	$f(B_1) = q$
Disequilibrium determinant	$D = f(A_1B_1)\, f(A_2B_2) - f(A_1B_2)\, f(A_2B_1)$	
Recombination fraction	c	
Diploid population size	N	
Generation number	t	

Fitness differences are assumed to be caused only by differential viability from conception to mating, and the general model is given in Table 1.

Table 1. Relative fitnesses

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$1-r_1-s_1+e_{11}$	$1-r_1$	$1-r_1-s_2+e_{12}$
A_1A_2	$1-s_1$	1	$1-s_2$
A_2A_2	$1-r_2-s_1+e_{21}$	$1-r_2$	$1-r_2-s_2+e_{22}$

Several models are of particular interest:

Additive	$e_{ij}=0$	$i,j=1,2$
Multiplicative	$e_{ij}=r_is_j$	
Completely symmetric	$r_i=s_i=s$, $e_{ij}=e$	

Thus in a completely symmetric model all double homozygotes have the same fitness and all single homozygotes have the same fitness. These are 1-2s and 1-s, respectively, in the additive case, and $(1-s)^2$ and 1-s in the multiplicative case.

Populations were assumed to have N monecious breeding individuals every generation among which random mating together with random selfing occured. The numerical analysis was performed by transition probability matrix methods where possible, otherwise by Monte Carlo simulation. In a diploid model the state of the population is descibed by the numbers, N_i, $i = 1, \ldots, 10$, of individuals of each genotype. For example N_1 can refer to individuals of genotype A_1B_1/A_1B_1 and so on, and $\sum_{i=1}^{10} N_i = N$. The expected frequencies of the gametes yielded by this population after recombination are computed, and from these the zygote frequencies after random mating. The expected frequencies of surviving individuals are computed using the fitnesses and from these frequencies the N individuals in the next generation were sampled from the multinominal distribution. In the Monte Carlo program (written by Dr. Joseph Felsenstein) a single sample was drawn using N uniform pseudo-random numbers and in the transition probability matrix method the probability of all possible states was computed. With $N = 4$ there are 715 possible states, but with the *completely symmetric* model it was possible to lump these into 69 typical states, following rules of Kemeny and Snell

(1960), and so operate with a matrix of domension 69×69. Statistics of the population, such as probabilities of segregation after a specific number of generations were computed by using up to 1000 replicate runs of Monte Carlo simulation, or by repeated iteration of the transition probability matrix onto an appropriate vector (Hill 1969). With the completely symmetric model both Monte Carlo and transition matrix methods yield the same expected values for probabilities of segregation, for example, but there is no sampling error using the matrix.

The assumption will be made throughout that the populations are initially in linkage equilibrium, and that the gene frequencies are at the equilibrium value appropriate for infinite populations. Thus in the additive and multiplicative models $p_0 = r_2/(r_1 + r_2)$, $q_0 = s_2/(s_1 + s_2)$, $D_0 = 0$. This is an unstable position in the multiplicative model if $c < \left(\frac{r_1 r_2}{r_1 + r_2}\right)\left(\frac{s_1 s_2}{s_1 + s_2}\right)$ (Bodmer and Felsenstein 1967), but is useful for comparing the additive and multiplicative models. Thus the situation which is being simulated, at least for the additive model, is one where a large number of replicated sub-populations of finite size are drawn from an infinitely large population in equilibrium.

RESULTS AND DISCUSSION

With $N = 4$ and a completely symmetric model, the transition matrix method could be used, and results are given in Table 2. At generations 8 and 32 the probabilities that both loci A and B are still segregating, $P(A^sB^s)$, and the probability that locus A is still segregating, $P(A^s)$, are tabulated. $P(A^s)$ is a marginal probability and includes

Table 2. The completely symmetric model with $N = 4$ and additive (A) or multiplicative (M) fitnesses. Marginal, $P(A^s)$, and joint, $P(A^sB^s)$, probabilities of segregation are tabulated after t generations, and the dominant eigen value (λ) is given for the matrix of transitions between states in which both loci segregate

		Selective value(s)					
Recombination fraction		0	.125		.25		.5
			M	A	M	A	M
λ	½	.770	.816	.819	.864	.879	.953
	1/16	.826	.871	.874	.914	.923	.978
	0	.875	.925	.928	.965	.975	.998
$P(A^s)$							
$t = 8$	½	.449	.539	.547	.643	.683	.863
	1/16	.449	.551	.560	.670	.710	.895
	0	.449	.559	.568	.685	.727	.911
$t = 32$	½	.018	.045	.050	.108	.165	.480
	1/16	.018	.052	.057	.143	.206	.635
	0	.018	.068	.076	.229	.321	.820
$P(A^sB^s)$							
$t = 8$	½	.207	.295	.302	.417	.460	.745
	1/16	.245	.347	.354	.483	.523	.810
	0	.268	.380	.388	.525	.565	.841
$t = 32$	½	.000	.002	.003	.013	.021	.234
	1/16	.002	.012	.013	.054	.073	.470
	0	.009	.049	.054	.193	.264	.770

populations in which B is segregating and those in which B is fixed. Since A and B loci can be interchanged in the completely symmetric model $P(A^s) = P(B^s)$. Other probabilities, such as that of A segregating and B fixed, $P(A^sB^f)$, are readily obtained, since $P(A^sB^f) = P(A^s) - P(A^sB^s)$. Also included in Table 2 are numerical values of the dominant eigen value, λ, for the submatrix of transitions between states at which both loci are segregating. Since these are all transient states, $\lambda < 1$ (excluding models in which only the double heterozygotes are viable).

If both loci are neutral ($s = 0$) then $P(A^s)$ is independent of c, the recombination fraction, but $P(A^sB^s)$ increases, the tighter the linkage. In this case the probability of joint fixation is also increased; tightly linked loci tend either to segregate together or become fixed together and there is a deficiency of populations in which one locus is segregating and the other is fixed. Even with free recombination $P(A^sB^s) > P(A^s)P(B^s)$, since the loci are still not independent, for D only decreases on average to almost half its value in the preceding generation. With $s = 0$ and $c = 0$ the root $\lambda = 1\text{-}1/2N$, the same value as for a single neutral locus, for, eventually, populations still segregating at both loci comprise only two types of gamete, either A_1B_1 and A_2B_2 or A_1B_2 and A_2B_1. With $c = 0.5$, $\lambda > (1\text{-}1/2N)^2$, again indicating that the two loci are not fixed independently.

When the heterozygote is favoured at both loci, tight linkage increases both the marginal and joint probabilities of segregation. In the extreme case of the multiplicative model with $s = 0.5$ the eigen values show that at the steady state only 0.2% of populations segregating at both loci are fixed for at least one locus in a single generation if $c = 0$, whereas 4.7% are fixed at the steady state if $c = 0.5$. For $s = 0.125$ and $s = 0.25$ the multiplicative and additive models are compared in Table 2. There are seen to be quantitative differences in eigen values and probabilities of segregation between the models, particularly at the larger s value. However there do not appear to be any qualitative differences in the results, in contrast with the infinite population situation. Segregation is maintained longer with the additive model, for the double homozygote is less fit than in the multiplicative case. The additive model with $s = 0.5$ is not included, since the double homozygote is then inviable.

Populations of size only $N = 4$ are not of major interest, and it is important to know what relevance results obtained with this population size have to larger, yet finite, populations. Space does not permit a full discussion of this aspect here. However using a diffusion approximation of the form given by Hill and Robertson (1966) it can be shown that sufficient parameters to describe the changes in gene frequency distribution over the replicate populations are Nr_i, Ns_j and Ne_{ij}, for $i, j = 1, 2$, together with Nc and the initial gene frequencies and disequilibrium, where time is measured as a proportion of N. The diffusion theory is strictly only relevant to populations of large size but comparisons of results using different values of N, but the same Ns and Nc values have indicated that the approximations hold adequately for many descriptive purposes (Hill and Robertson 1966, 1968). A further test is given in Table 3 for the completely symmetric model with $Ns = 1$ and $Nc = 1/4$ for $N = 4$ (matrix method), $N = 8$ and $N = 16$ (Monte Carlo method with 500 replicates). There is reasonably good agreement except in the additive case with $N = 4$, in which the double homozygote is very unfit. The additive and multiplicative models agree more closely with the larger N values, as we

Table 3. Comparison of probabilities of segregation at different population size with completely symmetric model and constant $Ns = 1$ and $Nc = 1/4$. Matrix method for $N = 4$, Monte Carlo with 500 replicates for $N = 8$ and 16. $P(A^s)$ with Monte Carlo is computed as mean of marginal probabilities $P(A^s)$ and $P(B^s)$.

Model		generations × 1/N				
		N	1	2	4	8
Additive	$P(A^s)$	4	.879	.710	.472	.206
		8	.884	.683	.406	.148
		16	.895	.679	.390	.129
	$P(A^sB^s)$	4	.777	.523	.265	.073
		8	.784	.490	.218	.054
		16	.802	.500	.224	.042
Multiplicative	$P(A^s)$	4	.863	.670	.405	.143
		8	.862	.672	.386	.142
		16	.899	.682	.414	.137
	$P(A^sB^s)$	4	.753	.483	.226	.054
		8	.762	.500	.210	.056
		16	.820	.510	.216	.046

Table 4. Marginal probability of segregation at the A locus, $P(A^s)$, after 32 generations with $N=4$ for a multiplicative model in which the selective values are $r_1=r_2$ at the A locus and $s_1=s_2$ at the B locus. Four decimal places are shown for matrix results, 3 for Monte Carlo results with 1000 replicates.

Recombination fraction	s_1	r_1 .0	.125	.25	.5
1/2	.0	.0182	.040	.112	.468
	.125	.012	.0447	.111	.464
	.25	.012	.045	.1080	.483
	.5	.019	.047	.122	.4804
1/16	.0	.0182	.054	.106	.446
	.125	.024	.0515	.135	.520
	.25	.030	.072	.1427	.569
	.5	.064	.123	.225	.6348
0	.0	.0182	.034	.114	.454
	.125	.019	.0681	.157	.554
	.25	.064	.110	.2288	.616
	.5	.210	.360	.498	.8200

might anticipate, for with constant Ns and increasing N, the epistatic terms Ne became smaller. In fact in the diffusion approximation for the multiplicative model these terms in Ne, which are of order Ns^2, are ignored.

As a first departure from the completely symmetric model consider the model in which the pair of homozygotes at each locus have the same fitness, but the two loci are not identical, *i.e.* $r_1 = r_2 \neq s_1 = s_2$. A suitable way of studying this model is to consider the marginal probability of segregation $P(A^s)$ as affected by fitness differences at the A locus, the B locus and the recombination fraction. The case of $N = 4$ with multiplicative fitnesses was studied using Monte Carlo simulation with 1000 replicates and

results are given in Table 4. The standard error of a value of $P(A^s)$ is given from the binomial distribution as $0.032\{P(A^s)[1-P(A^s)]\}^{1/2}$. Also included in the table are values of $P(A^s)$ computed by matrix iteration from the symmetric model, and identified by an additional significant digit. With free recombination $P(A^s)$ is very little influenced by the effects at the B locus, but with tighter linkage, the larger the fitness differences at the B locus, the higher the probability of segregation at the A locus. For example, if $c=0$, A has a higher probability of segregation after 32 generations if it is neutral ($r_1=0.0$) with B having a large effect ($s_1=0.5$), than if A has effect ($r_1=0.25$) with B neutral ($s_1=0.0$).

Except where one locus is neutral it has not been found possible to develop a quantitative theory to predict the effects of linkage on probability of segregation. Some qualitative explanations are possible but space prohibits much discussion of these. Essentially it seems that random drift generates linkage disequilibrium between the loci, equally likely to be positive or negative, and this increases before there is much fixation (Hill and Robertson, 1968). With an excess of, say, coupling chromosomes there is an excess of A_1B_1/A_1B_1, A_1B_1/A_2B_2 and A_2B_2/A_2B_2 individuals each generation. Now with an additive completely symmetric model, with selective values s, these genotypes have relative fitnesses $1-2s$, 1 and $1-2s$, so at either locus the apparant heterozygote superiority approaches 2s, compared with s when there is no disequilibrium. An excess of repulsion chromosomes has the same effect. Sved (1968) describes this result in terms of the marginal fitness of the genotypes at one locus, say A_1A_1, by averaging over genotypes containing A_1A_1. For the model in which the equilibrium frequency is 0.5 at each locus the marginal fitnesses may be written in the additive model as follows:

$$A_1A_1 : 1-r-s\left[q^2+(1-q)^2+\frac{(1-2p)(1-2q)D+2D^2-(1-2q)D}{p^2}\right]$$

$$A_1A_2 : 1\quad -s\left[q^2+(1-q)^2-\frac{(1-2p)(1-2q)D+2D^2}{p(1-p)}\right]$$

$$A_2A_2 : 1-r-s\left[q^2+(1-q)^2+\frac{(1-2p)(1-2q)D+2D^2+(1-2q)D}{(1-p)^2}\right]$$

In the completely symmetric model r can be replaced by s. Clearly at $p=q=0.5$ the superiority of the heterozygote is enhanced by disequilibrium, since $D^2>0$. However, with departures from these frequencies this difference may be increased, for Hill and Robertson (1968) have shown that $E[(1-2p)(1-2q)D]>0$ and $E[(1-2q)D]=0$ if $D=0$ initially.

For a neutral gene (A) linked to one with strong heterozygote superiority (B) a detailed analysis will be given in a later paper, but a simple argument is possible. We can assume that B remains segregating for a long time in the population, so that there exist two subpopulations, one containing B_1 alleles, the other B_2 alleles. In the extreme case of no recombination there is no migration between these subpopulations. In the B_1 subpopulations there are two types of chromosome, A_1B_1 and A_2B_1 which are neutral relative to each other, so become fixed relatively rapidly. Similarly, in the B_2 subpopulation either A_1B_2 or A_2B_2 become fixed. With initial frequencies of 0.5 at the A locus and initial equilibrium it will follow that in half the populations the same allele will be fixed in the sub-populations. For example the only chromosomes remaining may be A_1B_1 and A_2B_2. In these subpopulations locus A will only fix at the same rate

as B, and this rate will be very slow when B is highly heterotic. Recombination effectively allows migration between the subpopulations so that they can not remain fixed for different alleles at the A locus. It is also possible to consider the case where A is heterotic but with smaller effects than B in this model, when fixation is merely retarded within the sub-population.

The results discussed so far have only included models in which the homozygotes have equal fitness (*i.e.* equilibrium frequency of 0.5). In these the average values of D over replicates E(D), remains zero if it is zero initially. However when the equilibrium frequency is not 0.5 at each locus, then $E(D) \neq 0$. A simple example is given in Table 5, for $N = 4$ and $r_1 = s_1 = 0.15$ and $r_2 = s_2 = 0.35$ in a multiplicative model. The

Table 5. Heterozygote superiority with homozygotes having unequal fitness. Multiplicative model with $r_1 = s_1 = 0.15$, $r_2 = s_2 = 0.35$, $N = 4$ and Monte Carlo simulation using 1000 replicates.

Generation	8			32		
Recombination fraction	1/2	1/16	0	1/2	1/16	0
$P(A^s)$	.470	.520	.512	.068	.070	.110
$P(A^sB^s)$	.213	.302	.311	.003	.019	.080
$E(D) \times 100$	.000	-.130	-.322	-.008	-.038	-.711

equilibrium frequencies for both A_1 and B_1 taken alone are 0.7. We see that linkage retards fixation, and that E(D) is negative with tight linkage. Within segregating lines we see that E(D) after 32 generations with $c = 0$ is $-.00711/.080 = -.08$, approximately. These values are highly significantly different from zero. A simple interpretation of these results is possible. In population in which there is an excess of coupling gametes the frequent genotypes A_1B_1/A_1B_1, A_1B_1/A_1B_2 and A_2B_2/A_2B_2 have fitnesses 0.7, 1.0 and 0.3, respectively, in the model of Table 5, and this becomes equivalent to selection in a population with equilibrium frequency 0.7. On the other hand when repulsion gametes are in excess the frequent phenotypes are A_1B_2/A_1B_2, A_1B_2/A_2B_1, A_2B_1/A_2B_1 with fitnesses 0.5, 1.0 and 0.5, respectively, and the equilibrium frequency is 0.5. Fixation will occur more rapidly in the coupling populations which have the more extreme equilibrium frequency (Robertson 1962), and there will be an excess of repulsion gametes in the segregating populations. This can occur in the case of complete dominance also (*i.e.* $r_1 = s_1 = 0$), and the excess of repulsion gametes leads to "pseudo-overdominance" (Comstock and Robinson 1952). Accompanying the pseudo-overdominance is a retardation of fixation.

SUMMARY

The effects of linkage on the probability of gene segregation in small populations are discussed. A model of two loci, each with two alleles is used, in which the individual loci are neutral or show heterozygote superiority for fitness and the pair of loci are combined either with additive or multiplicative models of fitness. Initially the populations are assumed to be in linkage and gene frequency equilibrium.

It is found that tighter linkage increases the probability of segregation after several generations of small population size. An increase is obtained in both the joint prob-

ability of segregation at the two loci and in the marginal probability of segregation at either individual locus, except when the other locus is neutral. If at both loci the two homozygotes do not have equal fitness an excess of repulsion chromosomes is found among segregating populations. Except with very small population size and large fitness differences the additive and multiplicative model give similar results, although they have different properties in infinite populations.

REFERENCES

Bodmer, W.F., and J. Felsenstein, 1967 Linkage and selection: theoretical analysis of the deterministic two locus random mating model. Genetics **57**: 232-265.

Bodmer W.F., and P.A. Parsons, 1962 Linkage and recombination in evolution. Advan. Genet. **11**: 1-100.

Comstock, R.E., and H.F. Robinson, 1952 Estimation of average dominance of genes. In "Heterosis", (J.W. Gowen, ed.) pp. 494-516. Iowa State College Press, Ames.

Hill, W.G., 1968 Population dynamics of linked genes in finite populations. Proc. XII Intern. Congr. Genet. **2**: 146-147.

Hill, W.G., 1969 On the theory of artificial selection in finite populations. Genet. Res. **13**: 143-163.

Hill W.G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8**: 269-294.

Hill, W.G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**: 226-231.

Jain, S.K., and R.W. Allard, 1966 The effects of linkage, epistasis and inbreeding on population changes under selection. Genetics **53**: 633-659.

Karlin, S., and J. McGregor, 1968 Rates and probabilities of fixation for two locus random mating finite populations without selection. Genetics **58**: 141-159.

Kemeny, J.G., and J.L. Snell, 1960 Finite Markov Chains. Van Nostrand, Princeton, N.J.

Latter, B.D.H., 1966 The interaction between effective population size and linkage. Genet. Res. **7**: 313-323.

Levin, B.R., 1968 Simulation of genetic systems. Proc. Intern. Conf. on Computer Appl. in Genet. Univ. Hawaii. Honolulu.

Lewontin, R.C., 1964 The interaction of selection and linkage. I. General considerations; optimum models. Genetics **49**: 49-67.

Lewontin, R.C., and K. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution **14**: 458-472.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. (in press).

Robertson, A., 1962 Selection for heterozygotes in small populations. Genetics **47**: 1291-1300.

Sved, J.A., 1968 The stability of linked loci with a small population size. Genetics **59**: 543-563.

25

Disequilibrium among several linked neutral genes in finite population

I. Mean changes in disequilibrium

by

William G. Hill

MORLEY JONES, R. 1960. Linkage distributions and epistacy in quantitative inheritance, *Heredity* **15**, 153–159.

ROBERTSON, A. 1952. The effect of inbreeding on the variation due to recessive genes, *Genetics* **37**, 189–207.

SCHNELL, F. W. 1961. Some general formulations of linkage effects in inbreeding, *Genetics* **46**, 947–957.

SERANT, D. AND VILLARD, M. 1972. Linearisation of crossing over and mutation in a finite random mating population, *Theor. Pop. Biol.* **3**, 249–257.

SLATKIN, M. 1972. On treating the chromosome as the unit of selection, *Genetics* **72**, 157–168.

STRICKBERGER, M. W. 1968. "Genetics," Macmillan, New York.

SVED, J. A. 1968. The stability of linked systems of loci with a small population size, *Genetics* **59**, 543–563.

SVED, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations, *Theor. Pop. Biol.* **2**, 125–141.

WATTERSON, G. A. 1970a. The effect of linkage in a finite random mating population, *Theor. Pop. Biol.* **1**, 72–87.

WATTERSON, G. A. 1970b. On the equivalence of random mating and random union of gametes models in finite, monecious populations, *Theor. Pop. Biol.* **1**, 233–250.

WRIGHT, S. 1933. Inbreeding and recombination, *Proc. Nat. Acad. Sci. U. S.* **19**, 420–433.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.

Asymptotically, the multilocus disequilibrium is always associated with disequilibrium among pairs or triplets of loci. With four loci, for example, δ^*_{ABCD} approaches $D_{AB}D_{CD}$ after a few generations if L_{AD} is fairly large and loci equally spaced; then if there is no two-locus disequilibrium there is no four-locus disequilibrium either. The argument applies in reverse: if there exists two-locus disequilibrium by chance alone at several pairs of loci one may also expect to find disequilibria involving more loci. Just as with two loci, the observation of multilocus disequilibrium is not necessarily due to selection in the way Franklin and Lewontin (1970) and Slatkin (1972) have discussed, for a recent line cross or immigration could be involved. In a subsequent paper we shall discuss the maintenance of multilocus disequilibrium among neutral genes by random sampling in finite population, where, though the mean is zero, the variance is not. Such problems have already been analysed for two loci (Hill and Robertson, 1968; Ohta and Kimura, 1969a; Sved, 1971).

Acknowledgment

I wish to thank Mrs. Jennifer Smith for ably programming and executing all the computations.

References

Bennett, J. H. 1954. On the theory of random mating, *Ann. Eugen.* **184**, 311–317.

Franklin, I. R. and Lewontin, R. C. 1970. Is the gene the unit of selection? *Genetics* **65**, 707–734.

Geiringer, H. 1944. On the probability theory of linkage in Mendelian heredity, *Ann. Math. Stat.* **15**, 25–57.

Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors, *J. Genet.* **8**, 299–309.

Hill, W. G. and Robertson, A. 1966. The effect of linkage on limits to artificial selection, *Genet. Res.* **8**, 269–294.

Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations, *Theor. Appl. Genet.* **38**, 226–231.

Karlin, S. and McGregor, J. 1968. Rates and probabilities of fixation for two locus random mating finite populations without selection, *Genetics* **50**, 141–159.

Kendall, M. G. and Stuart, A. 1969. "The Advanced Theory of Statistics," Vol. I. 3rd ed. Griffing, London.

Kimura, M. 1955. Random genetic drift in multi-allelic locus, *Evolution* **9**, 419–435.

Kimura, M. 1963. A probability method for treating inbreeding systems, especially with linked genes, *Biometrics* **19**, 1–17.

Ohta, T. and Kimura, M. 1969a. Linkage disequilibrium due to random genetic drift, *Genet. Res.* **13**, 47–55.

Ohta, T. and Kimura, M 1969b. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation, *Genetics* **63**, 229–238.

Michell, 1973. Random mating and random union of gamets models for finite dioecious populations, *Theor. Pop. Biol.* **4**, 222–247.

of $n \times$ total length ($L_{AB}, L_{AC}, \ldots, L_{AF}$) and in terms of L, which is $n \times$ distance between adjacent loci ($L = L_{AB} = \cdots = L_{EF}$).

Loci	Root (α_1)		Associated disequilibrium
2	$1 + L_{AB}$	$= 1 + L$	D_{AB}
3	$3 + L_{AC}$	$= 3 + 2L$	Δ_{ABC}
4	$2 + (2/3)L_{AD}$	$= 2 + 2L$	$D_{AB}D_{CD}$
5	$4 + (3/4)L_{AE}$	$= 4 + 3L$	$D_{AB}\Delta_{CDE}$ or $D_{DE}\Delta_{ABC}$
6	$3 + (3/5)L_{AF}$	$= 3 + 3L$	$D_{AB}D_{CD}D_{EF}$

We have seen that the root is associated with the combination of disequilibrium in which the minimum of recombination occurs. With six loci this implies three adjacent pairs of two locus disequilibria, for its magnitude is affected by recombination between only three pairs of adjacent loci. Any other disequilibrium, such as $D_{AB}D_{CD}D_{EF}$, is the sum of the rates of each of its components, i.e., $3(1 + L)$ as seen above. It is reasonable to infer that matrices such as $\mathbf{M}_4$ to $\mathbf{M}_6$ can be constructed for seven or more loci, and that for seven loci, for example, the relevant terms will be a seven locus disequilibrium, 21 of the form $D_{AB}\partial_{CDEFG}$, 35 of the form $\Delta_{ABC}\delta_{DEFG}$ and 105 of the form $D_{AB}D_{CD}\Delta_{EFG}$, and that the total disequilibrium for seven loci will equal the sum of these. Of its components, the ones with least recombination will be $D_{AB}D_{CD}\Delta_{EFG}, D_{AB}\Delta_{CDE}D_{FG}, \Delta_{ABC}D_{DE}D_{FG}$, for which the root will be $2(1 + L) + (3 + 2L) = 5 + 4L$. For eight loci the relevant term will be $D_{AB}D_{CD}D_{EF}D_{GH}$ and the root $4(1 + L)$. Thus we suggest that, with equal spacing of m loci and sufficiently large values of L, the asymptotic rates of approach to equilibrium of the m-locus total disequilibrium are:

$$m \text{ even: } \alpha_1 = (m/2)(1 + L)$$
$$m \text{ odd: } \alpha_1 = [(m - 3)/2](1 + L) + (3 + 2L) = 1 + [(m + 1)/2](1 + L),$$

these corresponding to the number of pairs of loci if m is even, or number of pairs, plus a triplet of loci, if m is odd. If m is even and with $D^*_{m(t)}$ defining the expected value of any m locus total disequilibrium at generation t, then

$$D^*_{m(t+T)} \sim D^*_{m(T)}e^{-m(1+L)t/2n}$$

so long as t is sufficiently large for the asymptotic rates to be relevant. If the loci are unequally spaced the relevant root can be obtained by considering the possible arrangements of two and three locus disequilibria which minimize the total amount of recombination.

Our suggestion is, therefore, that the asymptotic rate of breakdown of multilocus disequilibrium is roughly proportional to the number of loci. This contrasts with the rate of loss of k multiple alleles at a single locus which is proportional to $k(k - 1)$ (Kimura, 1955).

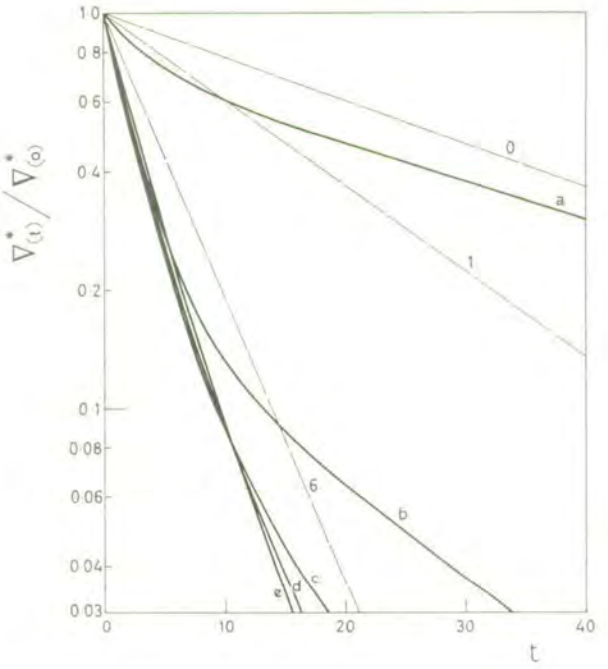

FIG. 6. As Fig. 5, but for six loci with $\nabla^*_{(t)}/\nabla^*_{(0)}$ for alternative values of (L_{AB}, L_{BC}, L_{CD}, L_{DE}, L_{EF}) as follows: a: (0 0 0 0 0), b: (0 6 0 0 0), c: (0 3 0 3 0), d: (0 0 6 0 0), e: ($\frac{6}{5}$ $\frac{6}{5}$ $\frac{6}{5}$ $\frac{6}{5}$ $\frac{6}{5}$).

higher initial rate of breakdown than if the middle pair are tightly linked, but the asymptotic rate of breakdown is much slower, as expected from Fig. 2.

Effects are rather less clear cut with six loci (Fig. 6). Before the alternative configurations have much influence on the asymptotic rate of breakdown of ∇^* most of the initial disequilibrium has been lost. The observed regularity of change in ∇^* hides large oscillations in its component terms ∇ and $D_{AB}D_{CD}D_{EF}$, etc.

8. DISCUSSION

The analysis has been restricted to only six loci, but some speculation about results for more seems justified. Consider the case of equally spaced loci, where the relevant parameter of population size $\times$ chromosome map length is sufficiently large that it is linearly related to the smallest root, α_1 (Figs. 2–4). For two to six loci the values of α_1 and the associated disequilibrium component which remains segregating asymptotically and constitutes almost all of the total disequilibrium are summarised below. The value of α_1 is expressed both in terms

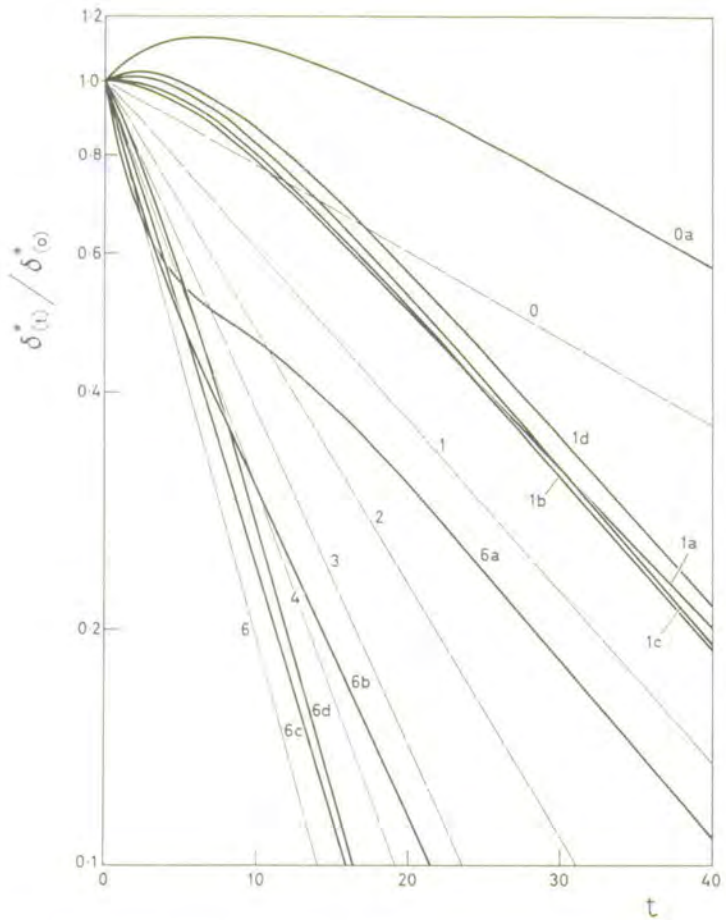

FIG. 5. Change in four-locus total disequilibrium (solid lines) for a population started from an inbred line cross. The disequilibrium is expressed as a proportion of its initial value ($\delta^*_{(t)}/\delta^*_{(0)}$) for alternative values of the parameters (L_{AB}, L_{BC}, L_{CD}) as follows: a: (0 1 0) L_{AD}, b: ($\frac{1}{6}$ $\frac{2}{3}$ $\frac{1}{6}$) L_{AD}, c: ($\frac{1}{3}$ $\frac{1}{3}$ $\frac{1}{3}$) L_{AD}, d: ($\frac{1}{2}$ 0 $\frac{1}{2}$) L_{AD}, and the coefficient is the value of L_{AD}. For example 6b implies (L_{AB}, L_{BC}, L_{CD}) = (1 4 1). The proportional change in two-locus disequilibrium ($D_{(t)}/D_{(0)}$) (dashed lines) is also shown for alternative values of L_{AB}.

(e.g. $L_{AD} = 1$), the value δ^* depends little on the configuration of the individual loci. but although the asymptotic rate of breakdown of δ^* is similar to that for a pair of loci, A, D separated by the same distance L_{AD}, the four locus disequilibrium is higher at any generations. With larger values of L_{AD} (e.g. $L_{AD} = 6$ in Fig. 5) the pattern of breakdown of disequilibrium depends much more on the configuration of the loci. If the outer pairs are tightly linked there is a slightly

two homozygous lines, such that all segregating genes have an initial frequency of 0.5 and all pairs of genes are in disequilibrium. In the F1 of a cross the genotypes at single loci are not in Hardy–Weinberg equilibrium, and the rate of breakdown of disequilibrium between the F1 and F2 is higher than in subsequent generations: in infinite population D_{AB} declines by $2[a \mid b]$ rather than $[a \mid b]$ subsequently, but we shall ignore this difference, which is not important for genes that are tightly linked.

In a cross of two homozygous lines the initial values of the disequilibrium are given below.

Two loci: $D_{AB} = \pm 1/4$

Three loci: $\Delta_{ABC} = 0$. For example, if the initial configuration in the two loci of the cross is $ABC/A'B'C'$, $r_{ABC} = q_{AB} = q_{AC} = q_{BC} = p_A = p_B = p_C = 1/2$, and $\Delta_{ABC} = 0$ from (3).

Four loci: There are alternative types of initial configuration. If there are an even number of alleles in the measured disequilibrium coming from one line (e.g. ABCD/A'B'C'D' or ABC'D'/A'B'CD), then $\delta_{ABCD} = -1/8$, $D_{AB}D_{CD} = D_{AC}D_{BD} = D_{BC}D_{AD} = 1/16$ and the total disequilibrium (8) of Slatkin (1972) is $\delta^*_{ABCD} = 1/16$. If there are an odd number of alleles in the disequilibrium from one line (e.g. ABCD'/A'B'C'D), all the above quantities have the same magnitude but opposite sign.

Five loci: $\partial_{ABCDE} = D_{AB}\Delta_{CDE} = \cdots = D_{DE}\Delta_{ABC} = 0$ for all configurations.

Six loci: With an even number of alleles in the specified disequilibrium coming from the same line (e.g. ABCDEF/A'B'C'D'E'F'), $\nabla_{ABCDEF} = 1/4$, $D_{AB}\delta_{CDEF} = \cdots = -1/32$, $\Delta_{ABC}\Delta_{DEF} = \cdots = 0$, $D_{AB}D_{CD}D_{EF} = \cdots = 1/8$ and $\nabla^*_{ABCDE} = 53/32$. The signs are reversed if an odd number of alleles come from one line.

Examples of the pattern of breakdown of four- and six-locus disequilibrium in populations derived from a line cross are given in Figs. 5 and 6. In each case the total disequilibrium (i.e., δ^* or ∇^*) is shown as a proportion of its value at generation 0, and for comparison rates of breakdown of two locus disequilibrium are shown on the same graphs. A population of size $n = 40$ was used, but similar results would be obtained if the same L values were used and time was expressed in terms of generations/n. Since the ordinates of Figs. 5 and 6 are plotted on a logarithmic scale, the decline in disequilibrium at two loci and the asymptotic declines with four or six loci are linear.

With very tight linkage ($L_{AD} \to 0$) there is initially an increase in δ^*_{ABCD}; the absolute values of its components δ_{ABCD} and $D_{AB}D_{CD}$ etc. are all declining, but they are of opposite sign and δ changes most quickly. When L_{AD} is small

$L_{BC} > 0, L_{DE} > 0$, then $\alpha_1 = 3 + L_{AB} + L_{EF}$ corresponding also to $D_{AB}D_{CD}D_{EF}$. If $L_{BC} = 0$ and $L_{CD} = 0$ four diagonal elements of $\mathbf{H}_5$ are zero, and it turns out that the smallest root of the matrix corresponding to these is $2 + L_{AB} + L_{EF}$. But when $L_{BC} = L_{DE} = 0$ the only zero term of $\mathbf{H}_5$ corresponds to $\Delta_{ABC}\Delta_{DEF}$ and the smallest root is $6 + L_{AB} + L_{EF}$. Other values can be derived from these and the list of relevant elements above. In all cases considered so far the smallest root has been real; but when $L_{BC} = L_{DE} = 0$, α_1 is imaginary over the range $3.78 < L_{CD} < 5.75$ approximately. Thus in Fig. 4 the real part over this range has been plotted. The imaginary part of the two conjugate roots reaches a value of 0.43 in the middle of the range.

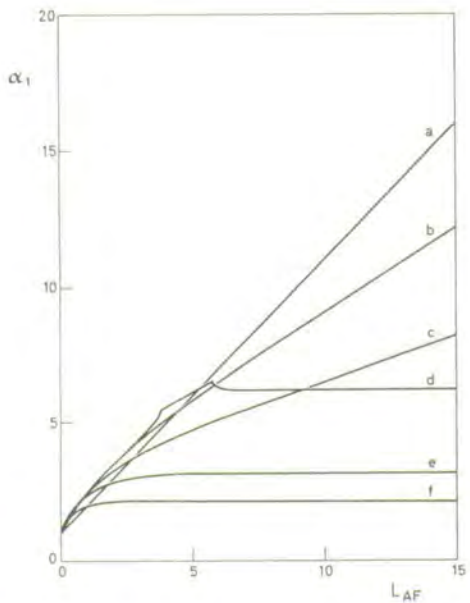

FIG. 4. Smallest root (α_1) of the matrix for six loci, $\mathbf{P}_6$, expressed as a function of L_{AF} for alternative spacing of genes. The values of $(L_{AB}, L_{BC}, L_{CD}, L_{DE}, L_{EF})$ are given by a: $(\frac{1}{2}\ 0\ 0\ 0\ \frac{1}{2})L_{AF}$, b: $(\frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5})L_{AE}$, c: $(0\ \frac{1}{3}\ \frac{1}{3}\ \frac{1}{3}\ 0)L_{AE}$, d: $(0\ 0\ 1\ 0\ 0)L_{AE}$, e: $(0\ \frac{1}{2}\ 0\ \frac{1}{2}\ 0)L_{AE}$, f: $(0\ 1\ 0\ 0\ 0)L_{AE}$.

7. Populations Based on Crosses of Inbred Lines

The rate of breakdown of linkage disequilibrium is most likely to be of interest in populations formed from a recent cross or following some immigration when considerable disequilibrium is likely. We illustrate some of the results with the most extreme and easily defined example, populations derived from a cross of

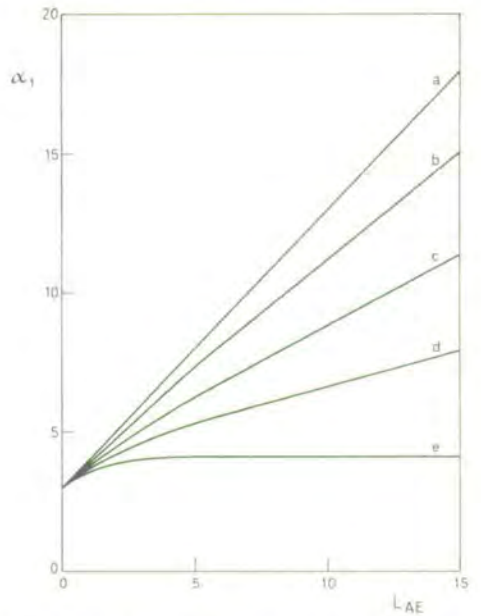

FIG. 3. Smallest root (α_1) of the matrix for five loci, $\mathbf{P}_5$, expressed as a function of L_{AE} for alternative spacing of genes. The values of (L_{AB}, L_{BC}, L_{CD}, L_{DE}) are given by a: $(\frac{1}{2}\ 0\ 0\ \frac{1}{2})L_{AE}$, b: $(\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4})L_{AE}$, c: $(0\ \frac{1}{2}\ \frac{1}{2}\ 0)L_{AE}$, d: $(0\ \frac{3}{4}\ \frac{1}{4}\ 0)L_{AE}$, e: $(0\ 1\ 0\ 0)L_{AE}$.

d. *Six Loci*

Inevitably, the results for six loci are more involved, as the results in Fig. 4 giving the smallest root with different configurations show. Some insight is obtained by considering those diagonal elements of $\mathbf{H}_6$ less than or equal to L_{AF} for all possible configurations. In all elements $L_{AB} + L_{EF}$ is present and excluded from the following list.

Term

$$D_{AB}\,\delta_{CDEF} \quad D_{EF}\,\delta_{ABCD} \quad \Delta_{ABC}\,\Delta_{DEF} \quad D_{AB}D_{CD}D_{EF} \quad D_{AB}D_{CE}D_{DF} \quad D_{AC}D_{BD}D_{EF}$$
$$D_{AB}D_{CF}D_{DE} \quad D_{AD}D_{BC}D_{EF}$$

Element of

$\mathbf{G}_6$	8	8	6	3	3	3
$\mathbf{H}_6$	$L_{CD}+L_{DE}$	$L_{BC}+L_{CD}$	$L_{BC}+L_{DE}$	L_{CD}	$L_{CD}+2L_{DE}$	$2L_{BC}+L_{CD}$

Thus we obtain the following results as $L_{BE} = L_{BC} + L_{CD} + L_{DE}$ becomes large. If $L_{BC} = L_{CD} = L_{DE}$ the smallest root is that corresponding to $D_{AB}D_{CD}D_{EF}$, so $\alpha_1 = 3 + L_{AB} + L_{EF} + L_{BC}$; and if all five loci are equally spaced so that $L_{AB} = \cdots = L_{EF} = L_{AF}/5$, then $\alpha_1 = 3 + (3/5)L_{AF}$. If $L_{CD} = 0$ and both

outside pairs of genes, and in addition the following terms, identified by the component of $\mathbf{w}_5$:

Term					
∂_{ABCDE}	$D_{AB}\,\Delta_{CDE}$	$D_{AC}\,\Delta_{BDE}$	$D_{AD}\,\Delta_{BCE}$	$D_{AE}\,\Delta_{BCD}$	$D_{BC}\,\Delta_{ADE}$
Element of $\mathbf{G}_5$					
$L_{BC}+L_{CD}$	L_{CD}	$2L_{BC}+L_{CD}$	$2L_{BC}+2L_{CD}$	$2L_{BC}+2L_{CD}$	$2L_{BC}+L_{CD}$
Term					
$D_{BD}\,\Delta_{ACE}$	$D_{BE}\,\Delta_{ACD}$	$D_{CD}\,\Delta_{ABE}$	$D_{CE}\,\Delta_{ABD}$	$D_{DE}\,\Delta_{ABC}$	
Element of $\mathbf{G}_5$					
$2L_{BC}+2L_{CD}$	$2L_{BC}+2L_{CD}$	$L_{BC}+2L_{CD}$	$L_{BC}+2L_{CD}$	L_{BC}	

The corresponding diagonal elements of $\mathbf{G}_5$ are 15 for ∂_{ABCDE} and 4 for each of the product terms (from Eq. (21)).

The latent roots, α, have the form

$$\alpha = L_{AB} + L_{DE} + \text{funct}(L_{BC}, L_{CD})$$

in which L_{BC} and L_{CD} are interchangeable variables. The smallest root is plotted against L_{AE} ($n \times$ total length) in Fig. 3 for different configurations. As L_{AE} becomes large, it is seen that α_i is linear in L_{AE}, with the slope dependent on the configuration, but can be simply expressed as

$$\alpha_1 \sim 4 + L_{AB} + L_{DE} + \min(L_{BC}, L_{CD}),$$

where $\min(L_{BC}, L_{CD})$ equals the smaller of L_{BC} and L_{CD}. For example, if $L_{BC} = L_{AE}$, then $\alpha_1 \sim 4$, so long as L_{AE} exceeds 6 approximately. With equally spaced loci, i.e. $L_{AB} = \cdots = L_{DE} = L_{AE}/4$, $\alpha_1 \sim 4 + (3/4)L_{AE}$.

The terms of $\mathbf{H}_5$ listed above help to explain these results. The smallest diagonal element of $\mathbf{G}_5 + \mathbf{H}_5$ is seen to equal $4 + L_{AB} + L_{DE} + \min(L_{BD}, L_{CD})$, associated with either $D_{DE}\,\Delta_{ABC}$ or $D_{AB}\,\Delta_{CDE}$, and either or both of these elements will be much smaller than other diagonal elements if $L_{BC} + L_{CD}$ is large. Also it can be seen from (21b) that the two off-diagonal elements of $\mathbf{G}_5$, giving $D_{DE}\,\Delta_{ABC}$ in terms of $D_{AB}\,\Delta_{CDE}$, and vice versa, are zero. Other off-diagonal elements are zero, or much smaller than $L_{BC} + L_{CD}$, if it is large. Thus we will find that the smallest eigenvalue of $\mathbf{G}_5 + \mathbf{H}_5$ is given by the smallest diagonal element, which will be associated with an eigenvector comprising zeros in all elements except that corresponding to the appropriate diagonal element. For example, if L_{CD} is small and L_{BC} is large, the smallest root approaches $4 + L_{AB} + L_{DE} + L_{CD}$, and the associated right eigenvector approaches $(0\ 1\ 0\ \ldots\ 0)'$; i.e., there is maintenance of $D_{AB}\,\Delta_{CDE}$ alone.

If L_{BC} becomes large, the roots become

$$\alpha \sim 2 + L_{AB} + L_{CD}\,, \qquad \alpha \sim 7 + L_{AB} + L_{CD} + L_{BC}\,,$$
$$\alpha \sim 1 + L_{AB} + L_{CD} + 2L_{BC}\,, \qquad \alpha = 3 + L_{AB} + L_{CD} + 2L_{BC}\,,$$

which can also be written

$$\alpha \sim 2 + L_{BC}\,, \quad \alpha \sim 7 + L_{AD}\,, \quad \alpha \sim 1 + L_{AD} + L_{BC}\,, \quad \alpha = 3 + L_{AD} + L_{BC}\,.$$

Space does not permit a full discussion of the associated eigenvectors of $\mathbf{P}_4$, nor, for five or six loci, analysis of other than the smallest root. Thus for comparison the smallest root, α_1, is plotted in Fig. 2 against L_{AD} for different configurations within the chromosome. This graph can be obtained directly from Fig. 1, but shows more clearly how much the asymptotic rate of breakdown of four locus disequilibrium depends on the gene arrangement. If loci are equally spaced, $L_{BC} = L_{AD}/3$ and $\alpha_1 \sim 2 + (2/3)L_{AD}$ if L_{AD} is large.

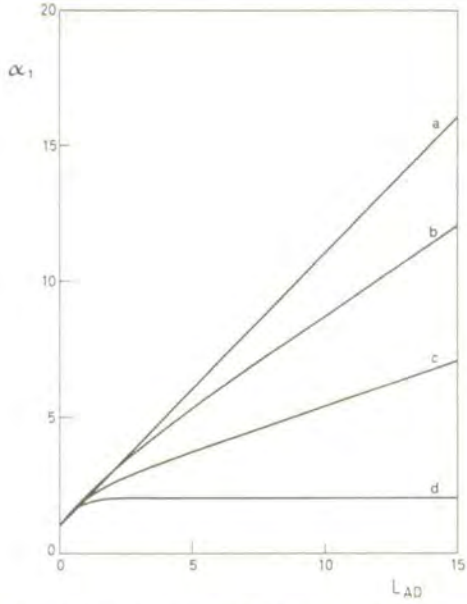

FIG. 2. Smallest root (α_1) of the matrix for four loci, $\mathbf{P}_4$, expressed as a function of L_{AD} for alternative spacing of genes. The values of (L_{AB}, L_{BC}, L_{CD}) are given by a: $(\frac{1}{2}\ 0\ \frac{1}{2})L_{AD}$, b: $(\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3})L_{AD}$, c: $(\frac{1}{6}\ \frac{2}{3}\ \frac{1}{6})L_{AD}$, d: $(0\ 1\ 0)L_{AD}$.

c. *Five Loci*

With five loci it is useful to list the diagonal elements of the recombination matrix $\mathbf{H}_5$. Each contains $L_{AB} + L_{DE}$, which is $n \times$ sum of lengths between the

if the two end pairs A, B and C, D are close together the rate of decline of disequilibrium is slow. Further analysis of $\mathbf{P}_4$ reveals that as L_{BC} becomes very large, the right eigenvector associated with the root $2 + L_{AB} + L_{CD}$ approaches $(0\ 1\ 0\ 0)'$; i.e., the combination which remains segregating is $D_{AB}D_{CD}$, which is not changed by recombination between B and C. The term D_{AB} declines at a rate of $1 + L_{AB}$ and D_{CD} at a rate of $1 + L_{CD}$, giving an overall rate of $2 + L_{AB} + L_{CD}$ as we find. Since the total disequilibrium δ^*_{ABCD} (Eq. (8)) is the sum of the four terms δ_{ABCD}, $D_{AB}D_{CD}$, $D_{AC}D_{BD}$ and $D_{AD}D_{BC}$ and does not specify an eigenvector of $\mathbf{P}_4$, its asymptotic rate of breakdown is also given by the smallest value of α.

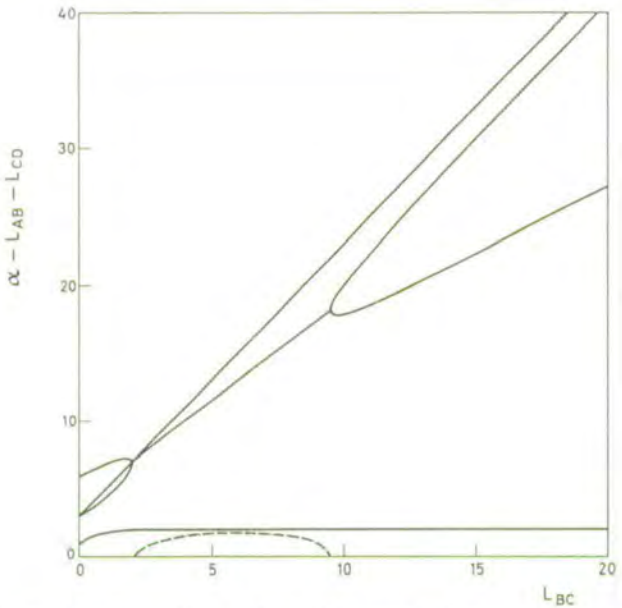

FIG. 1. Latent roots ($\alpha - L_{AB} - L_{CD}$) of the matrix for four loci, $\mathbf{P}_4$, expressed as a function of L_{BC}. The absolute value of the imaginary part of two conjugate roots is shown by a dotted line.

Figure 1 also shows that two of the roots are imaginary over the range $2 < L_{BC} < 9.5077$, but since these are not the smallest roots they are unlikely to induce much oscillation into the system after a few generations. If L_{BC} is very small, the roots from (35) and (36) can be shown to be

$$\alpha = 1 + L_{AB} + L_{CD} + 7L_{BC}/5 + O(L_{BC}^2),$$
$$\alpha = 3 + L_{AB} + L_{CD} + 2L_{BC}/3 + O(L_{BC}^2),$$
$$\alpha = 3 + L_{AB} + L_{CD} + 2L_{BC},$$
$$\alpha = 6 + L_{AB} + L_{CD} + 14L_{BC}/15 + O(L_{BC}^2).$$

If mutation is included in the model (Section 5b), $\mathbf{P}$ has to be multiplied by the scalar matrix $\Pi_m(1 - \eta_A - \epsilon_A)\mathbf{I}$ for m loci. Thus the eigenvalues λ, of $\mathbf{P}$ are multiplied by the factor $\Pi_m(1 - \eta_A - \epsilon_A)$ and, if η_A and ϵ_A are of order n^{-1}, the roots α are increased by $\sum_m(n\eta_A + n\epsilon_A)$ if n is large.

6. Rates of Approach to Equilibrium

We shall now use the results obtained in the earlier sections to study the rates of approach to linkage equilibrium with different numbers of loci. For simplicity we shall assume that the population is sufficiently large that the diffusion-type results obtained in Section 5 can be adopted.

a. *Two and Three Loci*

The pattern of approach to linkage equilibrium with two loci is given by (11). In terms of the chromosome length and assuming n to be reasonably large, the rate is $\alpha = 1 + L_{AB}$ per n generations, where $L_{AB} = nl_{AB}$. For three loci, we have from (12) that the rate is $\alpha = 3 + L_{AC}$. Thus we have

$$D_{AB(t)} = D_{AB(0)}e^{-(1+L_{AB})t/n},$$
$$\Delta_{ABC(t)} = \Delta_{ABC(t)} = \Delta_{ABC(0)}e^{-(3+L_{AC})t/n}.$$

b. *Four Loci*

With four loci we have to consider the four roots of $\mathbf{P}_4$, and some of the necessary analysis has been outlined in Section 5. In particular, we find that these roots are a function only of $L_{AB} + L_{CD}$ and L_{BC}. Also the matrix $\mathbf{H}_4$ (30) can be expressed as

$$\mathbf{H}_4 = (L_{AB} + L_{CD})\mathbf{I} + L_{BC}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

so the latent roots must be of the form $\alpha = L_{AB} + L_{CD} + \text{funct}(L_{BC})$. The characteristic equation (34) reduces to the following two equations,

$$\alpha = 3 + L_{AB} + L_{CD} + 2L_{BC} \tag{35}$$

and

$$(\alpha - L_{AB} - L_{CD})^3 - (\alpha - L_{AB} - L_{CD})^2(3L_{BC} + 10) + (\alpha - L_{AB} - L_{CD})(2L_{BC}^2 + 21L_{BC} + 27) - (4L_{BC}^2 + 32L_{BC} + 18) = 0. \tag{36}$$

Equation (36) can be solved in the usual way for cubics.

The solutions for the latent roots for four loci are given in Fig. 1, in which $\alpha - L_{AB} - L_{CD}$ is plotted against L_{BC}. The smallest root never exceeds $2.06 + L_{AB} + L_{CD}$, so that no matter the total map length of the chromosome,

yet they are quantitatively different. To try and minimise confusion we shall refer to λ as an eigenvalue and α as a (latent) root, or rate. Usually we consider α only if n is large. Then the values of α corresponding to $\lambda = (1 - 1/n),...,$ $(1 - 1/n)(...)(1 - 5/n)$ are 1, 3, 6, 10 and 15, which are roots of $\mathbf{M}_4$, $\mathbf{M}_5$ and $\mathbf{M}_6$.

From (33), the solutions for α with recombination are

$$|\, n\mathbf{P} - n\mathbf{I} + \alpha\mathbf{I}\,| = 0.$$

Letting $n \to \infty$ and substituting (32), we obtain the solutions

$$|\, \mathbf{G} + \mathbf{H} - \alpha\mathbf{I}\,| = 0, \tag{34}$$

so that for given values of L (i.e., L_{AB}, L_{AC}, etc.) the solutions for α, and also the latent vectors corresponding to real roots, do not depend on n so long as it is large.

An example of the utility of the asymptotic results is given in Table 1 for the four locus case. Two models are given: $L_{AB} = L_{BC} = L_{CD} = 1$ and $L_{AB} = 0$, $L_{BC} = 1, L_{CD} = 2$ for a range of values of n. For small n, the values of $[a \mid b]$, $[a \mid c]$, etc. were found from the inverse of (29), λ's were computed directly from (33) and α's as $\alpha = n(1 - \lambda)$, while Eq. (34) was used for $n \to \infty$. The latent roots given in Table 1 show changes of not more than 20% as n is increased from 20 towards infinity. The ratio of the smallest to the largest root at each level of n shows much less change. Thus the asymptotic results are good approximations unless n is very small. We notice in Eq. (30) that, for large n, $\mathbf{H}$ and the roots are functions of $L_{AB} + L_{CD}$, rather than either separately. This is illustrated in Table 1, and it is seen that, even with small n the alternative configurations of $L_{AB} = 1$ and $L_{AB} = 0$ with $L_{AB} + L_{CD} = 2$ make little difference to the roots.

TABLE 1

An Example for Four Loci of the Effect of Changes of Population Size (n) on Latent Roots (α)

	$L_{AB} = L_{BC} = L_{CD} = 1$				$L_{AB} = 0, L_{BC} = 1, L_{CD} = 2$			
n	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
20	3.441	5.342	5.842	7.375	3.406	5.324	5.824	7.339
40	3.625	5.799	6.382	8.075	3.606	5.788	6.371	8.053
80	3.722	6.051	6.680	8.455	3.712	6.045	6.674	8.442
160	3.772	6.184	6.837	8.653	3.767	6.181	6.834	8.646
320	3.797	6.252	6.918	8.754	3.795	6.251	6.916	8.750
640	3.810	6.287	6.959	8.805	3.809	6.286	6.958	8.803
1280	3.816	6.304	6.979	8.830	3.816	6.304	6.979	8.829
2560	3.820	6.314	6.991	8.844	3.820	6.314	6.990	8.843
$\to\infty$	3.823	6.322	7.000	8.856	3.823	6.322	7.000	8.856

and similar expressions can be written for the other elements of $\mathbf{R}_4$. With several loci it is more meaningful to work with chromosome map lengths l, where

$$l_{AB} = -(1/2)\log(1 - 2[a \mid b]), \tag{29}$$

which have the additive property,

$$l_{AC} = l_{AB} + l_{BC}$$

without making any assumptions of small recombination fractions (Haldane, 1919). When the latter assumption holds, we have from (29)

$$l_{AB} = [a \mid b] + O([a \mid b]^2).$$

Thus (28) becomes

$$[abcd] = 1 - l_{AB} - l_{BC} - l_{CD} + O(n^{-2}).$$

Letting $L_{AB} = nl_{AB}$, etc., which are terms of order unity, we obtain, by similar expansion of the other elements of $\mathbf{R}_4$,

$$\mathbf{R}_4 = \mathbf{I}_4 - n^{-1} \times \begin{pmatrix} L_{AB} + L_{BC} + L_{CD} & 0 & 0 & 0 \\ 0 & L_{AB} + L_{CD} & 0 & 0 \\ 0 & 0 & L_{AB} + 2L_{BC} + L_{CD} & 0 \\ 0 & 0 & 0 & L_{AB} + 2L_{BC} + L_{CD} \end{pmatrix} + O(n^{-2}). \tag{30}$$

The same method can be used for the matrices $\mathbf{R}_5$ and $\mathbf{R}_6$, and we write

$$\mathbf{R} = \mathbf{I} - n^{-1}\mathbf{H} + O(n^{-2}) \tag{31}$$

where $\mathbf{H}_4$ is given by (30). From (27) and (31)

$$\begin{aligned} \mathbf{P} &= \mathbf{MR} \\ &= \mathbf{I} - n^{-1}(\mathbf{G} + \mathbf{H}) + O(n^{-2}). \end{aligned} \tag{32}$$

The eigenvalues λ, of $\mathbf{P}$ are given by the solutions to

$$|\mathbf{P} - \lambda\mathbf{I}| = 0. \tag{33}$$

Now consider the transformed eigenvalues $\alpha = n(1 - \lambda)$, which correspond to the roots of an equivalent diffusion equation, perhaps with change of sign, with α being the rate of change per n generations. The term eigenvalue is commonly applied to both matrices and their corresponding differential equations,

If there are five or more loci with interference, the formulae for the disequilibria given by Bennett (1954) contain terms absent from $\mathbf{w}_5$ or $\mathbf{w}_6$. For example, the expression $\hat{\sigma}^0_{ABCDE}$ contains terms such as $p_A D_{BC} D_{DE}$. To take account of the interference it therefore becomes necessary to return to the matrices $\mathbf{L}_5$ and $\mathbf{U}_5$, which are of dimension 52×52 so that the calculation involved is much greater.

There is much evidence of the existence of interference is diploid species (e.g.(Strickberger, 1968, p. 333)). However its magnitude is significant only among genes which are very tightly linked, when the double crossover frequency is reduced by a large proportion below its already low value. In subsequent sections we shall be concerned with crossover frequencies between adjacent loci of order n^{-1}, for these are values at which both drift and recombination have appreciable effects on the rate of breakdown of disequilibrium. Unless n is very small, in chromosome regions with adjacent genes having crossover frequencies of about n^{-1} or less between adjacent genes, almost all crossovers in any generation will be singles even without interference. Therefore, interference which reduces double crossover frequencies is unlikely to have any important effect and we do not consider interference further.

5. Asymptotic Methods for Large Population Size

The matrices $\mathbf{P}$ (i.e., $\mathbf{P}_4 = \mathbf{M}_4\mathbf{R}_4, \ldots, \mathbf{P}_6 = \mathbf{M}_6\mathbf{R}_6$) and their eigenvalues and vectors depend on population size, but some generality can be achieved by considering these matrices as n becomes large, but with product terms, such as $n[a \mid b]$, of population size $\times$ crossover probabilities remaining finite. Therefore $[a \mid b]$ has order n^{-1}, and we make some of the same assumptions as are used in constructing the corresponding diffusion equations (Ohta and Kimura, 1969a). Where necessary, we illustrate the methods with $\mathbf{M}_4$, $\mathbf{R}_4$ and $\mathbf{P}_4$, given by (18), (19) and (20); if the matrices are not subscripted, the results are general.

The moment-generating matrices for drift can be written, for large n, as

$$\mathbf{M} = \mathbf{I} - n^{-1}\mathbf{G} + O(n^{-2}) \tag{27}$$

with, for example,

$$\mathbf{G}_4 = \begin{pmatrix} 7 & 2 & 2 & 2 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 2 & -1 \\ -1 & -1 & -1 & 2 \end{pmatrix}$$

from (18). Let us assume the loci A, B, C, and D are arranged in that order on the chromosome. The first diagonal element of $\mathbf{R}_4$ is, if all recombination fractions are small,

$$\begin{aligned}[abcd] &= (1 - [a \mid b])(1 - [b \mid c])(1 - [c \mid d]) \\ &= 1 - [a \mid b] - [b \mid c] - [c \mid d] + O(n^{-2})\end{aligned} \tag{28}$$

The result (27) is given by Ohta and Kimura (1969b) and Serant and Villard (1972), but the method has been outlined since it can readily be extended to three or more loci. The change in disequilibria among m loci turns out to be simply the product over all m loci of the probability of no recombinants at each. For example,

$$\delta_{ABCD(t+1)} = \delta_{ABCD(t)}\, \Pi_4(1 - \eta_A - \epsilon_A). \tag{28}$$

The other terms in the matrix $\mathbf{M}_4$, such as $D_{AB}D_{CD}$, change by the same proportion. Thus mutation can be incorporated into the changes in expected disequilibria in finite population as a scalar matrix with, for four loci for example, elements $\Pi_4(1 - \eta_A - \epsilon_A)$. Because the matrix is scalar the order of multiplication of this matrix with $\mathbf{M}_4\mathbf{R}_4$ is immaterial.

In the subsequent sections we shall again assume that there is no mutation, unless stated to the contrary.

c. *Interference*

It is possible to include interference in the analysis of the four locus haploid model. Following Bennett (1954) let us define a measure of four point interference by

$$g_{AB.CD} = [ab \mid cd]/([ab][cd] - [abcd])$$

and, if there is no interference, $g_{AB.CD} = 1$. Bennett obtains a four locus disequilibrium δ^0_{ABCD} such that, even with interference, $\delta^0_{ABCD(t)} = [abcd]\, \delta^0_{ABCD(t-1)}$ in infinite population. It can be shown that

$$\begin{aligned}\delta^0_{ABCD} = \delta_{ABCD} &+ (1 - g_{AB.CD})\, D_{AB}D_{CD} + (1 - g_{AC.BD})\, D_{AC}D_{BD} \\ &+ (1 - g_{AD.BC})\, D_{AD}D_{BC}\,.\end{aligned}$$

Thus we define a matrix $\mathbf{K}_4$ of full rank,

$$\mathbf{K}_4 = \begin{pmatrix} 1 & 1 - g_{AB.CD} & 1 - g_{AC.BD} & 1 - g_{AD.BC} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and a vector $\mathbf{w}^0_{4(t)}$ with the same elements as $\mathbf{w}_{4(t)}$, but δ^0_{ABCD} instead of δ_{ABCD}. Then

$$\begin{aligned}\mathbf{w}^0_{4(t)} &= \mathbf{K}_4\mathbf{w}_{4(t)} \\ &= \mathbf{K}_4\mathbf{M}_4\mathbf{R}_4\mathbf{K}_4^{-1}\mathbf{w}^0_{4(t-1)}\end{aligned}$$

from (20). The disequilibria at any generation can be evaluated from the starting conditions.

Comparison of (11) with (23) and of (12) with (25) shows that the haploid model is a good approximation when there is tight linkage, i.e., [*ab*] or [*abc*] close to unity, and n moderately large. If there is loose linkage any disequilibria declines so rapidly due to recombination that the finite study is not of interest. Following a suggestion of Sved (1971), it seems that a higher rate of breakdown in disequilibrium is predicted by the diploid model since there is nonreplacement sampling of chromosomes to form genotypes, leading to a slight excess of heterozygotes in which recombination can occur.

Equivalent diploid models for four or more loci have not been worked out, but it seems reasonable to infer from these two and three locus results that only small errors are incurred with the haploid model.

b. *Mutation*

The effects of recurrent mutation on the expected values of the disequilibria are easily computed in the haploid model, for we need operate only on chromosome frequencies between the stages of sampling each generation.

Let η_A be th probability of mutation of allele A to any other allele at the same locus in one generation, and let ϵ_A be the probability of mutation of any other allele to A. Let $p_{A'}$ be the frequency of all alleles other than A at that locus, $q_{A'B}$ be the frequency of chromosomes containing any of these alleles with allele B at the next locus and so on. Considering just mutation, frequencies in infinite population or expected frequencies in finite population at generation t in terms of those at $t-1$ (with subscripts suppressed for brevity) are, for example,

$$p_{A(t)} = p_A(1-\eta_A) + p_{A'}\epsilon_A$$
$$q_{AB(t)} = q_{AB}(1-\eta_A)(1-\eta_B) + q_{AB'}(1-\eta_A)\epsilon_B + q_{A'B}\epsilon_A(1-\eta_B) + q_{A'B'}\epsilon_A\epsilon_B .$$

Therefore

$$\begin{aligned} D_{AB(t)} &= q_{AB(t)} - p_{A(t)}p_{B(t)} \\ &= (q_{AB} - p_A p_B)(1-\eta_A)(1-\eta_B) + (q_{AB'} - p_A p_{B'})(1-\eta_A)\epsilon_B \\ &\quad + (q_{A'B} - p_{A'}p_B)\epsilon_A(1-\eta_B) + (q_{A'B'} - p_{A'}p_{B'})\epsilon_A\epsilon_B \\ &= D_{AB}(1-\eta_A)(1-\eta_B) + D_{AB'}(1-\eta_A)\epsilon_B + D_{A'B}\epsilon_A(1-\eta_B) \\ &\quad + D_{A'B'}\epsilon_A\epsilon_B \end{aligned} \tag{26}$$

But since $q_{AB} + q_{AB'} = p_A$ for example,

$$D_{AB} = -D_{AB'} = -D_{A'B} = D_{A'B'}$$

and Eq. (26) reduces to

$$\begin{aligned} D_{AB(t+1)} &= D_{AB}(1-\eta_A-\epsilon_A)(1-\eta_B-\epsilon_B) \\ &= D_{AB(t)}\,\Pi_2(1-\eta_A-\epsilon_A) \end{aligned} \tag{27}$$

using the subscript to denote number of product terms and replacing "t".

It is clear that the analysis could be continued up to seven or more loci, but the labour involved in deriving $\mathbf{M}_7$, say, is prohibitive. Before studying the properties of matrices such as $\mathbf{P}_6 = \mathbf{M}_6\mathbf{R}_6$ further, we consider a few side issues ignored so far.

4. Extensions to the Analysis

a. *Diploid Models*

A diploid model of $N = n/2$ monecious individuals with random selfing can also be used; but with two alleles per locus there is multinomial sampling of N individuals from 10 possible genotypes with two loci or 36 possible genotypes with three loci and so on. The case of two loci has been evaluated in the diploid model (Kimura, 1963) and discussed by Watterson (1970a). In Watterson's (1970b) terminology the model used by Kimura (1963) and here, for two and three loci, is one of random mating of zygotes, with mating occurring by independent trials. In our terminology Kimura's result becomes

$$\mathbf{v}_{2(t)} = \begin{pmatrix} [ab] & [a \mid b] \\ 1/n & 1 - 1/n \end{pmatrix} \mathbf{v}_{2(t-1)} ,$$

$$D_{AB(t)} = ([ab] - 1/n)\, D_{AB(t-1)} \tag{23}$$

(cf. (11)). With three loci, a straightforward but lengthy analysis yields the following relationship:

$$\begin{pmatrix} r_{ABC} \\ p_A q_{BC} \\ p_B q_{AC} \\ p_C q_{AB} \\ p_A p_B p_C \end{pmatrix}_{(t)}$$

$$= \begin{pmatrix} [abc] & [a \mid bc] & [b \mid ac] & [c \mid ab] & 0 \\ [bc]/n & (1 - 1/n)[bc] & [b \mid c]/n & [b \mid c]/n & (1 - 2/n)[b \mid c] \\ [ac]/n & [a \mid c]/n & (1 - 1/n)[ac] & [a \mid c]/n & (1 - 2/n)[a \mid c] \\ [ab]/n & [a \mid b]/n & [a \mid b]/n & (1 - 1/n)[ab] & (1 - 2/n)[a \mid b] \\ 1/n^2 & (1 - 1/n)/n & (1 - 1/n)/n & (1 - 1/n)/n & (1 - 1/n)(1 - 2/n) \end{pmatrix} \mathbf{v}_{3(t-1)} . \tag{24}$$

It turns out that the vector $\mathbf{y}_3'$, which defines Δ_{ABC}, is also an eigenvector of the matrix in (24), and we obtain for the diploid model

$$\Delta_{ABC(t)} = (1 - 2/n)([abc] - 1/n)\, \Delta_{ABC(t-1)} . \tag{25}$$

For six loci, the vector $\mathbf{w}_6$ and matrices $\mathbf{M}_6$ and $\mathbf{R}_6$ have dimension 41. The terms of $\mathbf{w}_6$ are

$$\{\nabla_{ABCDEF},\ D_{AB}\delta_{CDEF}\ \text{... (15 terms)},\ \Delta_{ABC}\Delta_{DEF}\ \text{... (10 terms)},\ D_{AB}D_{CD}D_{EF}\ \text{... (15 terms)}\}.$$

Typical terms of $\mathbf{M}_6$ are given by:

$$\begin{aligned}\nabla_{ABCDEF(t)} = \frac{n-1}{n^5}[&(n^4-30n^3+150n^2-240n+120)\nabla_{ABCDEF}\\ &-2(7n^3-36n^2+36n)\Sigma_{15}D_{AB}\delta_{CDEF}-6(3n^3-14n^2+16n)\\ &\times\Sigma_{10}\Delta_{ABC}\Delta_{DEF}-4(n^3-6n^2)\Sigma_{15}D_{AB}D_{CD}D_{EF}]_{(t-1)},\end{aligned}\tag{22a}$$

$$\begin{aligned}(D_{AB}\delta_{CDEF})_{(t)} = \frac{n-1}{n^5}[&(n-1)(n^2-6n+6)\nabla_{ABCDEF}\\ &+(n-1)(n^3-6n^2+6n)D_{AB}\delta_{CDEF}\\ &+(n^3-6n^2+6n)\Sigma_8D_{AC}\delta_{BDEF}\\ &-2n(n-1)\Sigma_6D_{CD}\delta_{ABEF}+0\Sigma_4\Delta_{ABC}\Delta_{DEF}\\ &+n(n-2)(n-4)\Sigma_6\Delta_{ACD}\Delta_{BEF}\\ &-2n^2(n-1)\Sigma_3D_{AB}D_{CD}D_{EF}\\ &-2n^2\Sigma_{12}D_{AC}D_{BD}D_{EF}]_{(t-1)},\end{aligned}\tag{22b}$$

$$\begin{aligned}(\Delta_{ABC}\Delta_{DEF})_{(t)} = \frac{n-1}{n^5}[&(n-1)(n-2)^2\nabla_{ABCDEF}+0\Sigma_6D_{AB}\delta_{CDEF}\\ &+n(n-2)^2\Sigma_9D_{AD}\delta_{BCEF}+n(n-1)(n-2)^2\Delta_{ABC}\Delta_{DEF}\\ &+n(n-2)^2\Sigma_9\Delta_{ABD}\Delta_{CEF}+0\Sigma_9D_{AB}D_{CD}D_{EF}\\ &+n^2(n-2)\Sigma_6D_{AD}D_{BE}D_{CF}]_{(t-1)},\end{aligned}\tag{22c}$$

$$\begin{aligned}(D_{AB}D_{CD}D_{EF})_{(t)} = \frac{n-1}{n^5}[&(n-1)^2\nabla_{ABCDEF}+n(n-1)^2\Sigma_3D_{AB}\delta_{CDEF}\\ &+n(n-1)\Sigma_{12}D_{AC}\delta_{BDEF}+0\Sigma_6\Delta_{ABC}\Delta_{DEF}\\ &+n(n-2)\Sigma_4\Delta_{ACE}\Delta_{BDF}+n^2(n-1)^2D_{AB}D_{CD}D_{EF}\\ &+n^2(n-1)\Sigma_6D_{AB}D_{CE}D_{DF}+n^2\Sigma_8D_{AC}D_{BE}D_{DF}]_{(t-1)}.\end{aligned}\tag{22d}$$

The appropriate terms in the summation signs in (22a–d) can readily be deduced from the typical terms given and the total number of terms. The eigenvalues of $\mathbf{M}_6$ are $(1-1/n)$, $(1-1/n)(1-2/n)$ 10 times, $(1-1/n)(1-2/n)(1-3/n)$ 20 times, $(1-1/n)(1-2/n)(1-3/n)(1-4/n)$ 9 times and $(1-1/n)(1-2/n)(1-3/n)(1-4/n)(1-5/n)$. The elements of the diagonal matrix $\mathbf{R}_6$ are

$$\{[abcdef],\ [ab][cdef]\ \text{... (15 terms)},\ [abc][def]\ \text{... (10 terms)},\ [ab][cd][ef]\ \text{... (15 terms)}\}.$$

Then

$$\mathbf{w}_{4(t)} = \begin{pmatrix} \mathbf{y}_4' \\ \mathbf{y}_{41}' \\ \mathbf{y}_{42}' \\ \mathbf{y}_{43}' \end{pmatrix} \mathbf{v}_{4(t)}$$

which, after using (13), (15), (16) and (17), reduces to

$$\begin{aligned} \mathbf{w}_{4(t)} &= \mathbf{M}_4\mathbf{R}_4\mathbf{w}_{4(t-1)} \\ &= \mathbf{P}_4\mathbf{w}_{4(t-1)} \end{aligned} \tag{20}$$

where $\mathbf{P}_4 = \mathbf{M}_4\mathbf{R}_4$. Equation (20) can of course be used repeatedly with little computing expenditure, to obtain expected values of the disequilibria δ_{ABCD}, $D_{AB}D_{CD}$, $D_{AC}D_{BD}$, and $D_{AD}D_{BC}$ at generation t in terms of those at generation 0. The eigenvalues of $\mathbf{M}_4$ can be shown to be $(1 - 1/n)$, $(1 - 1/n)(1 - 2/n)$, $(1 - 1/n)(1 - 2/n)$ and $(1 - 1/n)(1 - 2/n)(1 - 3/n)$, but an explicit form for the eigenvalues of $\mathbf{M}_4\mathbf{R}_4$ has not been obtained.

The same approach can be used to derive matrices $\mathbf{M}_5$ and $\mathbf{M}_6$ for five and six loci, respectively. With five loci the vector $\mathbf{w}_{5(t)}$ and matrices $\mathbf{R}_5$ and $\mathbf{M}_5$ have dimension 11, corresponding to the terms $\{\partial_{ABCDE}, D_{AB}\Delta_{CDE}, \ldots, D_{DE}\Delta_{ABC}\}$. The transformations that define $\mathbf{M}_5$ can be expressed for the following typical terms.

$$\begin{aligned} \partial_{ABCDE(t)} &= \frac{(n-1)(n-2)}{n^4} \\ &\quad \times [(n^2 - 12n + 12)\,\partial_{ABCDE} - 6n\Sigma_{10}D_{AB}\Delta_{CDE}]_{(t-1)}, \end{aligned} \tag{21a}$$

$$\begin{aligned} (D_{AB}\Delta_{CDE})_{(t)} &= \frac{(n-1)(n-2)}{n^4}[(n-1)\,\partial_{ABCDE} + n(n-1)\,D_{AB}\Delta_{CDE} \\ &\quad + n\Sigma_6 D_{AC}\Delta_{BDE} + 0\Sigma_3 D_{CD}\Delta_{ABE}]_{(t-1)} \end{aligned} \tag{21b}$$

where a subscript $(t - 1)$ is implied in all right-hand (r.h.s.) terms of (21a) and (21b). The six terms in (21b), typified by $D_{AC}\Delta_{BDE}$, are those in which one of the loci in D_{AB} (i.e., A or B) on the l.h.s. appear in D on the r.h.s. The other three terms have neither of the loci in D_{AB} in D on the r.h.s. There are, of course, 10 equations such as (21b), obtained by appropriate permutation of subscripts. The eigenvalues of $\mathbf{M}_5$ are found to be $(1 - 1/n)(1 - 2/n)$ five times; $(1 - 1/n)(1 - 2/n)(1 - 3/n)$ five times, and $(1 - 1/n)(1 - 2/n)(1 - 3/n)(1 - 4/n)$. The diagonal matrix $\mathbf{R}_5$ has diagonal elements $([abcde], [ab][cde], \ldots, [de][abc])$.

$p_A(1-p_A)(1-2p_A)$, but eigenvectors for fourth moments involves the population size n (e.g. Robertson, 1952). The analogy between $p_A(1-p_A)$ and D_{AB}, and between $p_A(1-p_A)(1-2p_A)$ and Δ_{ABC} will be demonstrated in a subsequent paper.

However we are still able to obtain a fairly simple result, as we illustrate for the case of four loci.

The elements of $\mathbf{v}_{4(t)}$ may be ordered

$$\mathbf{v}'_{4(t)} = (s_{ABCD}\,;\, p_A r_{BCD}\,,\, p_B r_{ACD}\,,\, p_C r_{ABD}\,,\, p_D r_{ABC}\,;\, q_{AB}q_{CD}\,,\, q_{AC}q_{BD}\,,\, q_{AD}q_{BC}\,;\, p_Ap_Bq_{CD}\,,\, p_Ap_Cq_{BD}\,,\, p_Ap_Dq_{BC}\,,\, p_Bq_{AD}\,,\, p_Bp_Dq_{AC}\,,\, p_Cp_Dq_{AB}\,;\, p_Ap_Bp_Cp_D)_{(t)} \tag{14}$$

so that $\mathbf{y}_4' = (1;\,-1\,-1\,-1\,-1;\,-1\,-1\,-1;\,2\,2\,2\,2\,2\,2;\,-6)$ from Eq. (4). Now we define three other vectors $\mathbf{y}'_{41}$, $\mathbf{y}'_{42}$ and $\mathbf{y}'_{43}$ such that $\mathbf{y}'_{41}\mathbf{v}_{4(t)} = (D_{AD}D_{CD})_{(t)}$, $\mathbf{y}'_{42}\mathbf{v}_{4(t)} = (D_{AC}D_{BD})_{(t)}$, and $\mathbf{y}'_{43}\mathbf{v}_{4(t)} = (D_{AD}D_{CD})_{(t)}$; for example

$$\mathbf{y}'_{41} = (0;\,0000;\,100;\,-10000-1;\,1).$$

To save space $\mathbf{L}_4$ and $\mathbf{U}_4$ are not tabulated, but it can be shown that

$$\mathbf{y}_4'\mathbf{L}_4 = [(n-1)/n^3][(n^2-6n+6)\,\mathbf{y}_4' - 2n(\mathbf{y}'_{41}+\mathbf{y}'_{42}+\mathbf{y}'_{43})], \tag{15}$$

$$\mathbf{y}'_{41}\mathbf{L}_4 = [(n-1)/n^3][(n-1)\,\mathbf{y}_4' + n(n-1)\,\mathbf{y}'_{41} + n(\mathbf{y}'_{42}+\mathbf{y}'_{43})], \tag{16}$$

and expressions for $\mathbf{y}'_{42}\mathbf{L}_4$, $\mathbf{y}'_{43}\mathbf{L}_4$ are given by appropriate permutation of the subscripts in (16). Thus these vectors form an invariant subspace of $\mathbf{L}_4$ and also of $\mathbf{L}_4\mathbf{U}_4$ for they are eigenvectors of $\mathbf{U}_4$:

$$\begin{aligned}&\mathbf{y}_4'\mathbf{U}_4 = [abcd]\,\mathbf{y}_4',\\ &\mathbf{y}'_{41}\mathbf{U}_4 = [ab][cd]\,\mathbf{y}'_{41}\,,\ \mathbf{y}'_{42}\mathbf{U}_4 = [ac][bd]\,\mathbf{y}'_{42}\,,\ \mathbf{y}'_{43}\mathbf{U}_4 = [ad][bc]\,\mathbf{y}'_{43}\,.\end{aligned} \tag{17}$$

Let us now define a vector $\mathbf{w}_{4(t)}$, with elements

$$\mathbf{w}'_{4(t)} = (\delta_{ABCD}\,,\, D_{AB}D_{CD}\,,\, D_{AC}D_{BD}\,,\, D_{AD}D_{BC})_{(t)}$$

(implying expected values as before) and matrices:

$$\mathbf{M}_4 = \frac{n-1}{n^3}\begin{pmatrix} n^2-6n+6 & -2n & -2n & -2n \\ n-1 & n^2-n & n & n \\ n-1 & n & n^2-n & n \\ n-1 & n & n & n^2-n \end{pmatrix}, \tag{18}$$

$$\mathbf{R}_4 = \begin{pmatrix} [abcd] & 0 & 0 & 0 \\ 0 & [ab][cd] & 0 & 0 \\ 0 & 0 & [ac][bd] & 0 \\ 0 & 0 & 0 & [ad][bc] \end{pmatrix}. \tag{19}$$

Noting that $D_{AB(t)} = q_{AB(t)} - (p_A p_B)_{(t)}$, we define the vector $\mathbf{y}_2' = (1 \ -1)$ so that $D_{AB(t)} = \mathbf{y}_2'\mathbf{v}_{2(t)}$. Since $\mathbf{y}_2'$ is a left eigenvector of both $\mathbf{U}_2$ and $\mathbf{L}_2$ associated with the eigenvalues $[ab]$ and $(1 - 1/n)$ respectively,

$$D_{AB(t)} = (1 - 1/n)[ab]\, D_{AB(t-1)} \tag{11}$$

in the haploid model, as is well known (Wright, 1933; Hill and Robertson, 1966).

For three loci, the relevant vector of moments is given by

$$\mathbf{v}_{3(t)}' = (r_{ABC}, \ p_A q_{BC}, \ p_B q_{AC}, \ p_C q_{AB}, \ p_A p_B p_C)_{(t)}$$

where the subscript on the vector implies that all terms have this subscript. It is then easy to show that for drift

$$\mathbf{L}_3 = \frac{1}{n^2}\begin{pmatrix} n^2 & 0 & 0 & 0 & 0 \\ n & n^2 - n & 0 & 0 & 0 \\ n & 0 & n^2 - n & 0 & 0 \\ n & 0 & 0 & n^2 - n & 0 \\ 1 & n-1 & n-1 & n-1 & (n-1)(n-2) \end{pmatrix},$$

and for recombination, using Bennett's (1954) results,

$$\mathbf{U}_3 = \begin{pmatrix} [abc] & [a \mid bc] & [b \mid ac] & [c \mid ab] & 0 \\ 0 & [bc] & 0 & 0 & [b \mid c] \\ 0 & 0 & [ac] & 0 & [a \mid c] \\ 0 & 0 & 0 & [ab] & [a \mid b] \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

As with two loci

$$\mathbf{v}_{3(t)} = \mathbf{L}_3\mathbf{U}_3\mathbf{v}_{3(t-1)}.$$

Now defining $\mathbf{y}_3' = (1 \ -1 \ -1 \ -1 \ 2)$, we have from (2),

$$\begin{aligned} \Delta_{ABC(t)} &= \mathbf{y}_3'\mathbf{v}_{3(t)} \\ &= (1 - 1/n)(1 - 2/n)[abc]\, \Delta_{ABC(t-1)}, \end{aligned} \tag{12}$$

since $\mathbf{y}_3'$ is a left eigenvector of $\mathbf{L}_3$ and $\mathbf{U}_3$.

The appropriate matrices, $\mathbf{L}_m$ and $\mathbf{U}_m$, for $m = 4, 5, 6$ (or more) loci are readily found, and have dimension 15, 52, and 203 respectively, where the terms are shown in equations (4), (5), and (6). But, the main problem is that the vectors $\mathbf{y}_m'$, which specify the disequilibria δ_{ABCD}, $\partial_{ABCD\dot{E}}^{*}$, ∇_{ABCDEF} (Eqs. (4)–(6)), are no longer eigenvectors of $\mathbf{L}_m$, the lower triangular matrix for drift. Thus we cannot simplify the general relation

$$\mathbf{v}_{m(t)} = \mathbf{L}_m\mathbf{U}_m\mathbf{v}_{m(t-1)}. \tag{13}$$

This problem is analogous with that of a single locus, where there is a simple eigenvector for second moments, $p_A(1 - p_A)$, and also for third moments,

Although it would be desirable to carry out the analysis using a full diploid model, this becomes too involved with more than three or so loci. Thus a haploid model is adopted in which a sample of n chromosome types are taken and their frequency distribution is squared to give expected genotype frequencies. From the genotype frequencies the expected gametic output after random mating with random selfing is computed, allowing for recombination. Then a sample of n gametes, or chromosome types, is sampled multinomially from the expected gamete frequencies. This completes one generation. The sampling and recombination are thus split into two successive stages, and the recombination occurs in a conceptually infinite population since expected frequencies are used. We shall analyse a full diploid model for simple cases in Section 4 and find our approximate haploid model to be adequate as far as it can be tested. This haploid model is a special case of that classified by Watterson (1970b, Sect. 3.1) as a "random union of gametes model."

Consider the case of two loci for which the results are well known, but which we shall review since the methods are illustrated most simply. Expected changes in D_{AB} can be expressed in terms of changes in q_{AB} and $p_A p_B$. These are obtained using the multinomial distribution, although regarding it as a bivariate binomial distribution (Kendall and Stuart, 1969, p. 141) and finding the expected values of moments such as $X_A X_B$, where X_A and X_B are the number of A and B genes in the sample of n gametes. The method is essentially that given by Serant and Villard (1972) and we do not give details. Extension to three or more loci is straightforward.

Since we always require expected values of quantities in finite population, we shall now use p_A, q_{AB}, r_{ABC}, D_{AB}, etc., with subscripts for generations if necessary, to denote expected values. Letting

$$\mathbf{v}_{2(t)} = \begin{pmatrix} q_{AB(t)} \\ (p_A p_B)_{(t)} \end{pmatrix}$$

we obtain

$$\mathbf{v}_{2(t)} = \mathbf{L}_2 \mathbf{U}_2 \mathbf{v}_{2(t-1)} . \tag{9}$$

In (9)

$$\mathbf{L}_2 = \begin{pmatrix} 1 & 1 \\ 1/n & 1 - 1/n \end{pmatrix} \tag{10}$$

and specifies the changes in the moments due to sampling; it is a moment-generating matrix of the type used by Robertson (1952) for single loci. Also in (9)

$$\mathbf{U}_2 = \begin{pmatrix} [ab] & [a \mid b] \\ 0 & 1 \end{pmatrix}$$

and expresses changes due to recombination, which precedes sampling (10) in our haploid model.

Although these measures of disequilibrium are most easily handled in our analysis, they do not seem to be the best operational definitions of multilocus disequilibria, which we take from Slatkin (1972). It is known that D_{AB} is the covariance of gene frequencies at two loci, which we can write as

$$D_{AB} = E[(x_A - p_A)(x_B - p_B)]$$

where x_A , $(x_B) = 1$ or 0 according to whether the A (B) allele is present or absent. Higher order disequilibria are defined in the same way and are here denoted by * to distinguish them from those of Bennett in Eqs. (4)–(6),

$$\delta^*_{ABCD} = E[(x_A - p_A)(x_B - p_B)(x_C - p_C)(x_D - p_D)].$$

On expanding and substituting $s_{ABCD} = E(x_A x_B x_C x_D)$, for example, we find

$$\begin{aligned} \delta^*_{ABCD} &= s_{ABCD} - \Sigma_4 p_A q_{BCD} + \Sigma_6 p_A p_B q_{CD} - 3p_A p_B p_C p_D \\ &= \delta_{ABCD} + D_{AB}D_{CD} + D_{AC}D_{BD} + D_{AD}D_{BC} \end{aligned} \tag{8}$$

using (1) and (4). Similarly, it can be shown that

$$\partial^*_{ABCDE} = \partial_{ABCDE} + \Sigma_{10} D_{AB} \Delta_{CDE}$$

$$\nabla^*_{ABCDEF} = \nabla_{ABCDEF} + \Sigma_{15} D_{AB} \delta_{CDEF} + \Sigma_{10} \Delta_{ABC} \Delta_{DEF} + \Sigma_{15} D_{AB} D_{CD} D_{EF}$$

and we also have that the total disequilibria for two and three loci are D_{AB} and Δ_{ABC} . Slatkin (1972) is incorrect in stating that his and Bennett's disequilibria are the same for more loci.

Apart from their simple definition, the disequilibria of Slatkin, which will be denoted *total* disequilibria, have the property that in a diploid population of size N the quantity for four loci, say,

$$2N\delta^{*2}_{ABCD}/[p_A(1 - p_A) \cdots p_D(1 - p_D)]$$

is the chi-square value to test for the total four locus disequilibrium and is equivalent to the test for third-order interaction in the analysis of variance. This property will be considered in more detail in a subsequent paper.

3. Basic Analysis

In the subsequent analysis there is assumed to be a finite random mating population with no selection or mutation, which comprises N diploid individuals or $n = 2N$ chromosomes.

which, in an infinitely large random mating population declines as

$$D_{AB(t)} = [ab]\, D_{AB(t-1)}$$

where $D_{AB(t)}$ is the magnitude of D_{AB} at generation t. With more loci we now choose as the disequilibria functions of the chromosome or gene frequencies such that they decline exponentially each generation in an infinite population. These functions have been given by Bennett (1954). For three loci he defines

$$\Delta_{ABC} = r_{ABC} - p_A D_{BC} - p_B D_{AC} - p_C D_{AB} - p_A p_B p_C$$

which, using (1), may be written

$$\Delta_{ABC} = r_{ABC} - p_A q_{BC} - p_B q_{AC} - p_C q_{AB} + 2p_A p_B p_C \,. \tag{2}$$

Because higher order disequilibria include so many terms we use a shorthand in which summation over all (say k) relevant combinations is denoted by Σ_k. Thus (2) becomes

$$\Delta_{ABC} = r_{ABC} - \Sigma_3 p_A q_{BC} + 2p_A p_B p_C \,. \tag{3}$$

With four or more loci the expression of the disequilibria is complicated by the possible existence of interference (Bennett, 1954). We shall simplify our presentation by assuming that there is no such interference, and defer the inclusion of interference to a subsidiary section of the paper. Then, for four loci, we have by rearrangement of Bennett's formula

$$\delta_{ABCD} = s_{ABCD} - \Sigma_4 p_A r_{BCD} - \Sigma_3 q_{AB} q_{CD} + 2\Sigma_6 p_A p_B q_{CD} - 6p_A p_B p_C p_D; \tag{4}$$

for five loci,

$$\begin{aligned} \partial_{ABCDE} = {} & t_{ABCDE} - \Sigma_5 p_A s_{BCDE} - \Sigma_{10} q_{AB} r_{CDE} + 2\Sigma_{10} p_A p_B r_{CDE} \\ & + 2\Sigma_{15} p_A q_{BC} q_{DE} - 6\Sigma_{10} p_A p_B p_C q_{DE} + 24 p_A p_B p_C p_D p_E; \end{aligned} \tag{5}$$

and for six loci,

$$\begin{aligned} \nabla_{ABCDEF} = {} & u_{ABCDEF} - \Sigma_6 p_A t_{BCDEF} - \Sigma_{15} q_{AB} s_{CDEF} - \Sigma_{10} r_{ABC} r_{DEF} \\ & + 2\Sigma_{15} p_A p_B s_{CDEF} + 2\Sigma_{60} p_A q_{BC} r_{DEF} + 2\Sigma_{15} q_{AB} q_{CD} q_{EF} \\ & - 6\Sigma_{20} p_A p_B p_C r_{DEF} - 6\Sigma_{45} p_A p_B q_{CD} q_{EF} + 24\Sigma_{15} p_A p_B p_C p_D q_{EF} \\ & - 120 p_A p_B p_C p_D p_E p_F \,. \end{aligned} \tag{6}$$

Thus there are 203 terms in the definition of ∇_{ABCDEF}. Bennett (1954) has shown that in infinite population

$$\begin{aligned} \Delta_{ABC(t)} &= [abc]\, \Delta_{ABC(t-1)}; & \delta_{ABCD(t)} &= [abcd]\, \delta_{ABCD(t-1)}; \\ \partial_{ABCDE(t)} &= [abcde]\, \partial_{ABCDE(t-1)}; & \nabla_{ABCDEF(t)} &= [abcdef]\, \nabla_{ABCDEF(t-1)} \,. \end{aligned} \tag{7}$$

and for predicting changes in chromosome frequencies and disequilibria when there is no selection in an infinitely large population (Bennett, 1954). However Bennett's results for many loci have not yet, apparently, been extended to finite populations. An attempt to do so is made in this paper. While the methods developed here can, in principle, be extended to more than six loci, this has not been attempted since the computations become too tedious. However, it is hoped that the results obtained will be sufficient to demonstrate the changes in disequilibria following, for example, a cross between two inbred populations.

This paper will be concerned solely with neutral genes. No claim is being made that this is the only case worth studying, but it is necessary to have an adequate theory for neutral genes before possible effects due to selection can be tested. As Bennett (1954) has demonstrated for infinite populations and we shall show in this paper for finite populations, the mean disequilibrium always tends to zero, but its rate of approach may be very slow if genes are tightly linked. Thus disequilibrium involving several loci which exists in a population may just be a consequence of the history of its foundation, rather than any existing selection effects. We shall deal in this paper only with mean values of disequilibria, but the extension to variances and covariances of disequilibria among neutral genes uses the methods adopted here and will appear in a subsequent paper. In that, it will be shown that, just as with two loci, considerable disequilibrium can arise.

2. Definitions

Consider a set of loci A, B,..., with any number of neutral alleles, and let p_A , p_B ,..., be the frequency of a specified allele at each. For these same alleles we also define the following typical chromosome frequencies q_{AB} , r_{ABC} , s_{ABCD} , t_{ABCDE} , and u_{ABCDEF} . We shall use the same notation for recombination as did Bennett (1954). The small letters *a*, *b*, *c*,..., represent the respective loci and a vertical bar separates the contributions from two homologous chromosomes. Thus $[ab \mid c]$ is the total frequency of all gamete types formed by a recombination between loci B and C (and A and C) but not between A and B, and $[abc]$ is the total frequency of nonrecombinants. For example, if an individual is a heterozygote ABC/A′B′C′ where A′, B′, C′ are alternative alleles, its chromosomal output will be:

$$\text{freq(ABC)} + \text{freq(A′B′C′)} = [abc],$$
$$\text{freq(ABC′)} + \text{freq(A′B′C)} = [ab \mid c].$$

There are obviously many ways in which the disequilibria between loci can be measured with two loci, that commonly used is

$$D_{AB} = q_{AB} - p_A p_B \qquad (1)$$

Reprinted from THEORETICAL POPULATION BIOLOGY
All Rights Reserved by Academic Press, New York and London

Vol. 5, No. 3, June 1974
Printed in Belgium

Disequilibrium Among Several Linked Neutral Genes in Finite Population

I. Mean Changes in Disequilibrium*

WILLIAM G. HILL

Statistical Laboratory, Iowa State University, Ames, Iowa 50010

and

Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN, Scotland†

Received May 1, 1973

Formulae are developed for computing changes in expected values in a finite population of linkage disequilibrium among neutral genes from more than two loci, although the exact analysis is taken up to only six loci. An essentially haploid model is used. As with two loci, the three-locus disequilibrium declines exponentially at all generations, but for $m > 3$ loci a matrix has to be constructed to give joint changes in the m-locus disequilibrium and products of disequilibria with fewer loci, for example of two $m/2$-locus disequilibria. The asymptotic rates of change in multilocus disequilibria depend on the arrangement of genes on the chromosome as well as its total length, but the initial rate of breakdown of disequilibrium from a line cross base is less dependent on the arrangement. With equally spaced loci the asymptotic rate of breakdown of m locus disequilibrium is roughly proportional to m. Although mutation and interference are excluded from the main analysis, it is shown how they can be incorporated.

1. INTRODUCTION

There is now an extensive literature on the prediction of the mean and variance of changes in linkage disequilibrium between pairs of neutral genes (e.g. Wright, 1933; Kimura, 1963; Hill and Robertson, 1966, 1968; Karlin and McGregor, 1968; Ohta and Kimura, 1969a, b; Sved, 1968, 1971; Watterson, 1970a; Michell, 1973). Also there already exists an adequate theory in population genetics for predicting the frequencies of different recombinant types from specified genotypes with many loci (Geiringer, 1944; Morley Jones, 1960; Schnell, 1961)

* Journal paper No. J-7541, Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project 1669. Supported in part by National Institutes of Health, Grant No. 13827.

† Permanent address.

26

Disequilibrium among several linked neutral genes in finite population

II. Variances and covariances of disequilibria

by

William G. Hill

three loci (Table 1), are due originally to Bennett (1954) and are convenient for use in population dynamics since explicit formulae can be given for the present values and for changes in disequilibria, both in infinite and finite populations. This specification of lack of association of gene frequencies, for example $\Delta_{ABC} = 0$, is not the same, however, as that usually defined for $2 \times 2 \times 2$ contingency tables and due originally to Bartlett (1935). Letting A′, B′ and C′ be alternative alleles to A, B and C, respectively, Bartlett's definition of no second-order association (corresponding to no three-locus disequilibrium) is

$$r_{ABC}r_{AB'C'}r_{A'BC'}r_{A'B'C} = r_{ABC'}r_{AB'C}r_{A'BC}r_{A'B'C'} .$$

This definition does not, however, lead to explicit expressions either for chromosome frequencies, or a three-locus disequilibrium. Thus, while Bartlett's criterion may be suitable for testing for association in data collected from the field, it does not lend itself to predictions of changes in frequencies as described in this paper.

Acknowledgment

I wish to thank Mrs. Jennifer Smith for carrying out the computing.

References

Bartlett, M. S. 1935. Contingency table interactions, *J. R. Statist. Soc. Suppl.* **2**, 248–252.

Bennett, J. H. 1954. On the theory of random mating, *Ann. Eugen.* **184**, 311–317.

Hill, W. G. 1974. Disequilibrium among several linked neutral genes in finite population. I. Mean changes in disequilibrium, *Theor. Pop. Biol.* **5**, 366–392.

Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations, *Theor. Appl. Genet.* **38**, 226–231.

Karlin, S. and McGregor, J. 1968. Rates and possibilities of fixation for two locus random mating finite populations without selection, *Genetics* **59**, 141–159.

Kimura, M. and Ohta, T. 1971. "Theoretical Aspects of Population Genetics," Princeton University Press, New Jersey, NJ.

Ohta, T. and Kimura, M. 1969. Linkage disequilibrium due to random genetic drift, *Genet. Res.* **13**, 47–55.

Rao, C. R. and Mitra, S. K. 1971. "Generalized Inverses," Wiley, New York.

Robertson, A. 1952. The effect of inbreeding on the variation due to recessive genes, *Genetics* **37**, 189–207.

Sved, J. A. 1968. The stability of linked systems of loci with a small population size, *Genetics* **59**, 543–563.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.

do not have an important effect $L_{AB} = L_{BC} = \frac{1}{2}L_{AC}$ is used. The scalar constants of 16 for $V(D_{AB})$ and 64 for $V(\Delta_{ABC})$ are the initial values of

$$1/p_A(1-p_A)\,p_B(1-p_B) \quad \text{and} \quad 1/p_A(1-p_A)\,p_B(1-p_B)\,p_C(1-p_C),$$

respectively.

	L_{AB}	0	$\frac{1}{2}$	2	8
max. 16 $V(D_{AB})$, t		0.162,35	0.126,26	0.086,17	0.049,8
	L_{AC}	0	$\frac{1}{2}$	2	8
max. 64 $V(\Delta_{ABC})$, t		0.291,11	0.239,7	0.159,6	0.075,2

With three loci, much higher values of $V(\Delta_{ABC})$ are achieved from a line cross than with initial equilibrium (cf. Fig. 1). Despite the initial disequilibrium, the pattern of change in $V(D_{AB})$ is not greatly different from that with initial equilibrium.

4. Discussion

The main purpose of this paper has been to illustrate methodology. Ohta and Kimura (1969) and others have been able to develop diffusion equation methods for finding the variance of the two-locus disequilibrium, which reduces to a recurrence relation in three moments, but it seems unlikely that this will be possible with three loci, where 16 moments are involved. The three-locus problem has been analysed here by the moment-generating matrix method, but it is improbable that more loci can be handled in this way, for although there seems to be no conceptual difficulty in setting up the necessary matrix, say for four loci involving eight moments, the labour involved in deriving it is likely to be prohibitive. It is hoped, however, that this and the preceding paper will draw attention to some of the problems involved with more than two loci with no selection. Previous studies seem to have been restricted to only two loci, yet since electrophoretic variants can be detected at several loci on a chromosome we need a theory which enables us to predict the behaviour of a population in the absence of selection if only as a basis for showing that selection takes place.

This analysis has been in terms of expectations taken over all replicate populations, but in many instances we are concerned solely with populations in which there is segregation at all the constituent loci of the disequilibrium. That problem will be considered in a further paper, for different analytical methods have to be used.

The definitions of disequilibria which are used here, for example Δ_{ABC} for

used, but the results would be essentially the same if a different value were taken. It is seen that $V(\Delta_{ABC})$ reaches a maximum more quickly than $V(D_{AB})$, and at a smaller value than the appropriate initial product of the variance of gene frequencies. The overall pattern is rather similar however. Notice that, while the smallest root of the $(A^2B^2C^2)$ matrix depends on the relative magnitude of the chromosome lengths (i.e., L_{AB}/L_{BC}), the pattern of change in $V(\Delta_{ABC})$ in the early generations depends mostly on the sum, $L_{AB} + L_{BC}$.

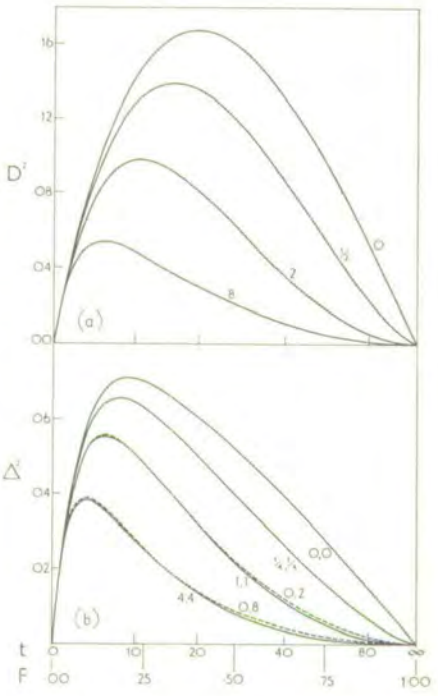

FIG. 1. $E(D^2)/p_A(1-p_A)\,p_B(1-p_B)$ and $E(\Delta^2)/p_A(1-p_A)\,p_B(1-p_B)\,p_C(1-p_C)$ shown as D^2 and Δ^2 in (a) and (b), respectively, plotted against time expressed as $F = 1-(1-1/n)^t$ and also against t, where t is generations, $n = 40$ is the haploid population size, and p_A, p_B, p_C are the gene frequencies in the initial population.

In populations started from a cross of two inbred lines, which provides an extreme case of disequilibrium, initially $D_{AB} = 1/4$ and $\Delta_{ABC} = 0$ (Hill, 1974). The variance of D_{AB} and Δ_{ABC} have been computed for a few examples of parameters. Both, of course, are initially zero, reach a maximum and gradually decline to zero. The times taken to reach the maximum t, and the values of the variances at the maximum are given below; in each case computations were done with $n = 40$. Since the relative magnitudes of L_{AB} and L_{BC} in the three-locus case

specifying (A^2B^2) and $(A^2B^2C^2)$, respectively. The two-locus case has been much studied already, and we include it here only for comparison. For three loci, the appropriate moment-generating matrix **Q**, has dimension 16 and its roots are given in Table 3. With recombination values of zero the smallest root α_1, of $\mathbf{P} = \mathbf{QR}$ is given in Table 4 as a function of L_{AB} and L_{BC}, where, for example,

TABLE 4

Smallest Root $\alpha_1 = \lim_{n\to\infty} n(1 - \lambda_1)$, of the Matrices Specified by (A^2B^2), (A^3B^3) as Functions of L_{AB}, and of $(A^2B^2C^2)$ as Functions of L_{AC} and Its Partition into L_{AB}, L_{BC}

Code		L_{AB}	0	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8	16	$\to\infty$
(A^2B^2)			1	1.225	1.400	1.628	1.824	1.933	1.978	1.994	2
(A^3B^3)			1	1.250	1.500	2.000	3.000	5.000	5.859	5.952	6
		L_{AC}	0	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8	16	$\to\infty$
	L_{AB}	L_{BC}									
$(A^2B^2C^2)$	$\frac{1}{2}L_{AC}$	$\frac{1}{2}L_{AC}$	1	1.237	1.449	1.798	2.248	2.633	2.855	2.951	3
	$\frac{3}{4}L_{AC}$	$\frac{1}{4}L_{AC}$	1	1.234	1.437	1.755	2.148	2.517	2.780	2.918	3
	L_{AC}	0	1	1.225	1.400	1.628	1.824	1.933	1.978	1.994	2

$L_{AB} = n \times$ map length of chromosome between A and B. (The roots $\alpha = \lim_{n\to\infty} n(1 - \lambda)$ of **P** are shown to be functions of L_{AB}, L_{BC}; etc., for equivalent matrices by Hill (1974)). With very tight linkage ($L_{AB} + L_{BC} \to 0$), the smallest root is that of **Q** and equals unity (see Table 3). With much recombination, the smallest root is 3 and corresponds to an eigenvector which has zero in all elements except that specifying the quantity $p_A(1 - p_A)\,p_B(1 - p_B)\,p_C(1 - p_C)$; this declines in magnitude at a rate three times as great as each of its constituent terms, such as $p_A(1 - p_A)$, because the genes at different loci are independent.

The magnitudes of $V(D_{AB})$ and $V(\Delta_{ABC})$ with initial equilibrium are compared in Fig. 1, where time is expressed in proportion to the inbreeding coefficient, $F = 1 - (1 - 1/n)^t$. In a population initially in equilibrium it can be seen from the appropriate row of Table 3 that the only quantity not initially zero in the vector **x** containing Δ^2_{ABC} is the product of frequencies

$$p_A(1 - p_A)\,p_B(1 - p_B)\,p_C(1 - p_C).$$

Thus the magnitude of Δ^2_{ABC} in subsequent generations is proportional to this product and the results in Fig. 1 are expressed as a ratio of $E(\Delta^2_{ABC})$ to the initial value of $p_A(1 - p_A)\,p_B(1 - p_B)\,p_C(1 - p_C)$. A haploid population size of 40 was

(Table 3 continued)

(A^2B^2CD)	20	1	5	9	4	1	$(1-2p_A)(1-2p_B)\,\delta_{ABCD}$, $D_{AB}\delta_{ABCD}$, $\Sigma_2(1-2p_A)\,D_{AB}\Delta_{BCD}$, $\Sigma_4(1-2p_A)\,D_{BC}\Delta_{ABD}$, $\Delta_{ABC}\Delta_{ABD}$, $\Sigma_2 D_{AB}D_{AC}D_{BD}$, $\Sigma_2 p_A(1-p_A)\,D_{BC}D_{BD}$, $\Sigma_2(1-2p_A)(1-2p_B)\,D_{AC}D_{BD}$, $D_{AB}^2D_{CD}$, $\Sigma_2 p_A(1-p_A)(1-2p_B)\,\Delta_{BCD}$, $(1-2p_A)(1-2p_B)\,D_{AB}D_{CD}$, $p_A(1-p_A)\,p_B(1-p_B)\,D_{CD}$
$(A^3B^2C^2)$	16	1	4	7	3	1	$(1-2p_A)(1-2p_B)(1-2p_C)\,\Delta_{ABC}$, $\Sigma_3(1-2p_A)\,D_{BC}\Delta_{ABC}$, $\Sigma_3(1-2p_A)(1-2p_B)\,D_{AC}D_{BC}$, Δ_{ABC}^2, $D_{AB}D_{AC}D_{BC}$, $\Sigma_3 p_A(1-p_A)(1-2p_B)(1-2p_C)\,D_{BC}$, $\Sigma_3 p_A(1-p_A)\,D_{BC}^2$, $p_A(1-p_A)\,p_B(1-p_B)\,p_C(1-p)_C$
(A^3BCD)	13	1	3	5	3	1	δ_{ABCD}, $p_A(1-p_A)\,\delta_{ABCD}$, $\Sigma_3(1-2p_A)\,D_{AB}\Delta_{ACD}$, $\Sigma_3 p_A(1-p_A)\,D_{AB}D_{CD}$, $p_A(1-p_A)(1-2p_A)\,\Delta_{BCD}$, $D_{AB}D_{AC}D_{AD}$, $\Sigma_3 D_{AB}D_{CD}$
(A^3B^2C)	10	1	2	4	2	1	$(1-2p_B)\,\Delta_{ABC}$, $D_{AB}D_{BC}$, $p_A(1-p_A)(1-2p_B)\,\Delta_{ABC}$, $(1-2p_A)(1-2p_B)\,D_{AB}D_{AC}$, $(1-2p_A)\,D_{AB}\Delta_{ABC}$, $D_{AB}^2D_{AC}$, $p_A(1-p_A)\,D_{AB}D_{BC}$, $p_B(1-p_B)\,D_{AC}$, $p_A(1-p_A)\,p_B(1-p_B)\,D_{AC}$, $p_A(1-p_A)(1-2p_A)(1-2p_B)\,D_{BC}$
(A^3B^3)	7	1	1	3	1	1	D_{AB}, $\Sigma_2 p_A(1-p_A)\,D_{AB}$, $p_A(1-p_A)\,p_B(1-p_B)\,D_{AB}$, $(1-2p_A)(1-2p_B)\,D_{AB}^2$, D_{AB}^3, $p_A(1-p_A)(1-2p_A)\,p_B(1-p_B)(1-2p_B)$
(A^4BC)	6	1	1	2	1	1	$(1-2p_A)\,\Delta_{ABC}$, $D_{AB}D_{AC}$, $p_A(1-p_A)(1-2p_A)\,\Delta_{ABC}$, $p_A(1-p_A)\,D_{AB}D_{AC}$, $p_A(1-p_A)\,D_{BC}$, $p_A^2(1-p_A)^2\,D_{BC}$
(A^4B^2)	6	1	1	2	1	1	$(1-2p_A)(1-2p_B)\,D_{AB}$, D_{AB}^2, $p_A(1-p_A)(1-2p_A)(1-2p_B)\,D_{AB}$, $p_A(1-p_A)\,D_{AB}^2$, $p_A(1-p_A)\,p_B(1-p_B)$, $p_A^2(1-p_A)^2\,p_B(1-p_B)$
(A^5B)	3	1	0	1	0	1	D_{AB}, $p_A(1-p_A)\,D_{AB}$, $p_A^2(1-p_A)^2\,D_{AB}$
(A^6)	3	1	0	1	0	1	$p_A(1-p_A)$, $p_A^2(1-p_A)^2$, $p_A^3(1-p_A)^3$

[a] List of all possible combinations, e.g. $\Sigma_3 D_{AB}D_{CD}$ denotes $D_{AB}D_{CD}$, $D_{AC}D_{BD}$, $D_{AD}D_{BC}$.

TABLE 3

Dimensions of Derived Matrices and their Roots, Together with the Terms Appearing in the Appropriate Vector of Moments

Order	Code	Dimension	Roots 1	3	6	10	15	Terms in vector of moments
4	(ABCD)	4	1	2	1			δ_{ABCD} , $\Sigma_3 D_{AB}D_{CD}$[a]
	(A^2BC)	3	1	1	1			$(1-2p_A)\,\Delta_{ABC}$, $D_{AB}D_{AC}$, $p_A(1-p_A)\,D_{BC}$
	(A^2B^2)	3	1	1	1			$(1-2p_A)(1-2p_B)\,D_{AB}$, D_{AB}^2 , $p_A(1-p_A)\,p_B(1-p_B)$
	(A^3B)	2	1	0	1			D_{AB} , $p_A(1-p_A)\,D_{AB}$
	(A^4)	2	1	0	1			$p_A(1-p_A)$, $p_A^2(1-p_A)^2$
5	(ABCDE)	11	0	5	5	1		∂_{ABCDE} , $\Sigma_{10} D_{AB}\Delta_{CDE}$
	(A^2BCD)	8	0	4	3	1		$(1-2p_A)\,\delta_{ABCD}$, $\Sigma_3 D_{AB}\Delta_{ACD}$, $p_A(1-p_A)\,\Delta_{BCD}$, $\Sigma_3 (1-2p_A)\,D_{AB}D_{CD}$
	(A^2B^2C)	6	0	3	2	1		$(1-2p_A)(1-2p_B)\,\Delta_{ABC}$, $D_{AB}\Delta_{ABC}$, $\Sigma_2 (1-2p_A)\,D_{AB}D_{BC}$, $\Sigma_2 (1-2p_A)\,p_B(1-p_B)\,D_{AC}$
	(A^3BC)	4	0	2	1	1		Δ_{ABC} , $p_A(1-p_A)\,\Delta_{ABC}$, $(1-2p_A)\,D_{AB}D_{AC}$, $p_A(1-p_A)(1-2p_A)\,D_{BC}$
	(A^3B^2)	4	0	2	1	1		$(1-2p_B)\,D_{AB}$, $p_A(1-p_A)(1-2p_B)\,D_{AB}$, $(1-2p_A)\,D_{AB}^2$, $p_A(1-p_A)(1-2p_A)\,p_B(1-p_B)$
	(A^4B)	2	0	1	0	1		$(1-2p_A)\,D_{AB}$, $p_A(1-p_A)(1-2p_A)\,D_{AB}$
	(A^5)	2	0	1	0	1		$p_A(1-p_A)(1-2p_A)$, $p_A^2(1-p_A)^2(1-2p_A)$
6	(ABCDEF)	41	1	10	20	9	1	∇_{ABCDEF} , $\Sigma_{15} D_{AB}\delta_{CDEF}$, $\Sigma_{10}\,\Delta_{ABC}\Delta_{DEF}$, $\Sigma_{15} D_{AB}D_{CD}D_{EF}$
	(A^2BCDE)	28	1	7	13	6	1	$(1-2p_A)\,\partial_{ABCDE}$, $\Sigma_4 D_{AB}\delta_{ACDE}$, $\Sigma_3\,\Delta_{ABC}\Delta_{ADE}$, $p_A(1-p_A)\,\delta_{BCDE}$, $\Sigma_6 (1-2p_A)\,D_{BC}\Delta_{ADE}$, $\Sigma_6 D_{AB}D_{AC}D_{DE}$, $\Sigma_4 (1-2p_A)\,D_{AB}\Delta_{CDE}$, $\Sigma_3 p_A(1-p_A)\,D_{BC}D_{DE}$

defines the reduction A = C, B = D from (ABCD) so that there are second moments at the A and B locus, and $\mathbf{x}$ is given by Eq. (11) and $\mathbf{Q}$ by (14), with each having dimension 3.

There are too many matrices for the elements of each to be given in full, but as a partial summary the eigenvalues are also given in Table 3. These are given as $\alpha = \lim_{n\to\infty} n(1 - \lambda)$, where λ is the eigenvalue, so that for the possible eigenvalues $\lambda = (1 - 1/n)$, $(1 - 1/n)(1 - 2/n)$, $(1 - 1/n)(1 - 2/n)(1 - 3/n)$,

$$(1 - 1/n)(1 - 2/n)(1 - 3/n)(1 - 4/n)$$

and

$$(1 - 1/n)(1 - 2/n)(1 - 3/n)(1 - 4/n)(1 - 5/n),$$

$\alpha = 1$, 3, 6, 10 and 15. The eigenvalues of $\mathbf{Q}$ are always a subset of those of the matrix $\mathbf{M}$, from which it is derived, and the existence of $\lim_{n\to\infty} n(1 - \lambda)$ for eigenvalues of $\mathbf{M}$ was demonstrated by Hill (1974). No simple algorithm for obtaining the multiplicity of the roots of $\mathbf{Q}$ from $\mathbf{M}$ has been found, and the values given in Table 3 were obtained by direct operations on the derived $\mathbf{Q}$ matrices. When there is no recombination, the roots of $\mathbf{Q}$ are also those of the product matrix $\mathbf{P} = \mathbf{QR}$, since $\mathbf{R}$ becomes an identity matrix.

The matrices $\mathbf{Q}$, are of triangular or block triangular form if the vector $\mathbf{x}$ contains moments of different order, for changes in moments of a specified order never involve one of higher order. An example is (A^4) where the moments are $p_A(1 - p_A)$ and $p_A^2(1 - p_A)^2$, and the matrix is given by (15).

b. *Covariance of Disequilibria*

We notice in Table 3 that for any specification involving first moments at one or more loci (e.g. (ABCD), (A^3B), (A^3B^2C)), all terms in the appropriate vector of moments are disequilibrium terms; contrast, for example, (A^3B) with (A^2B^2), where the latter includes the term $p_A(1 - p_A)\,p_B(1 - p_B)$. Thus if there is initial linkage equilibrium, all terms in the vector $\mathbf{x}_{(0)}$ are zero, and $\mathbf{x}_{(t)} = 0$ for all t. Now the crossproducts of two disequilibria (e.g. $D_{AB}D_{AC}$ and $D_{AB}\Delta_{ABC}$), appear in such vectors (e.g. (A^2BC) and (A^2B^2C), respectively). Thus if there is initial equilibrium, both the crossproducts of the disequilibria and their individual expected values are zero, and they are uncorrelated. If there is initial disequilibrium, but some recombination between the appropriate loci occurs, the covariances will approach zero as time increases. By contrast, terms such as $D_{AB}D_{AC}D_{BC}$ and $(1 - 2p_A)(1 - 2p_B)\,D_{AC}D_{BC}$ do not have expected values of zero, because they appear with $p_A(1 - p_A)\,p_B(1 - p_B)\,p_C(1 - p_C)$ in $(A^2B^2C^2)$.

c. *Variances of Disequilibria*

When there is initial linkage equilibrium, $V(D_{AB}) = E(D_{AB}^2)$ and $V(\Delta_{ABC}) = E(\Delta_{ABC}^2)$, and so the variances are obtained directly from iteration of the matrices

In each case

$$\mathbf{T} = \begin{pmatrix} 1 & -6 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

from Table 1, and $\mathbf{Q} = \mathbf{T}^*\mathbf{M}\mathbf{T}$ is given by

$$\mathbf{Q} = \frac{n-1}{n^3} \begin{pmatrix} n^2 & 0 \\ n-1 & n^2-5n+6 \end{pmatrix} \tag{15}$$

with eigenvalues $1 - 1/n$ and $(1 - 1/n)(1 - 2/n)(1 - 3/n)$. The single-locus result (i.e., A = B = C = D) can also be derived from the paper of Robertson (1952). It is possible, of course, to obtain the matrix for A = B = C = D by reduction of that for, say, A = B, C = D, rather than going right back to the basic matrix **M**.

The method can also be used for higher moments and the necessary elements of **M** are given for five and six loci by Hill (1974). Although it is always easy to specify the transformation matrices **T**, using Table 2, all of which have full column rank, the calculation $\mathbf{Q} = \mathbf{T}^*\mathbf{M}\mathbf{T}$ is very tedious by hand, especially for sixth moments where **M** has dimension 41×41. However if we write, for the $(k+1)$th moment

$$\mathbf{M} = \frac{1}{n^k} \sum_{i=0}^{k} n^i \mathbf{M}_{(i)}$$

such that $\mathbf{M}_{(i)}$ has the coefficients of n^i, we compute **Q** from

$$\mathbf{Q} = \frac{1}{n^k} \sum_{i=0}^{k} n^i \mathbf{T}^* \mathbf{M}_{(i)} \mathbf{T}. \tag{16}$$

The coefficient matrices $\mathbf{T}^*\mathbf{M}_{(i)}\mathbf{T}$ can be evaluated numerically on a computer and (16) used to find an algebraic formula for **Q** as a function of n.

3. Results

a. *Derived Vectors and Matrices*

The dimensions of the derived vectors **x** and matrices **Q**, and the moments contained in **x** are given in Table 3. A code is used to define the set of moments. For example, for moments of order 4, (ABCD) represents moments of order 1 at each locus, and so refers to the basic matrix **M**, and vector **w** (which are just special cases of the derived matrix and vector) of Hill (1974); whereas (A^2B^2)

Multiplying both sides by $\mathbf{T}^*$, we obtain

$$\begin{aligned}\mathbf{x}_{(t)} &= \mathbf{T}^*\mathbf{M}\mathbf{T}\mathbf{x}_{(t-1)} \\ &= \mathbf{Q}\mathbf{x}_{(t-1)}\,,\end{aligned} \tag{13}$$

where

$$\mathbf{Q} = \mathbf{T}^*\mathbf{M}\mathbf{T}.$$

In this example,

$$\mathbf{Q} = \frac{n-1}{n^3}\begin{pmatrix} n^2-4n+4 & 4n-8 & \cup \\ n-1 & n^2-2n+2 & n \\ n-1 & 2 & n^2-n \end{pmatrix} \tag{14}$$

which is the same matrix as given by Hill and Robertson (1968) for the case of no recombination, but with rows and columns permuted. The matrix $\mathbf{Q}$ is unique and has eigenvalues $1 - 1/n$, $(1 - 1/n)(1 - 2/n)$, $(1 - 1/n)(1 - 2/n)(1 - 3/n)$, a subset of those of $\mathbf{M}$.

In our haploid model (Hill, 1974) we have assumed that drift sampling follows recombination and that changes in disequilibria due to recombination are those appropriate for an infinite population. So for the vector $\mathbf{x}_{(t)}$, the relevant recombination matrix $\mathbf{R}$ is diagonal with elements $[ab]$, $[ab]^2$ and 1, associated with the elements $(1 - 2p_A)(1 - 2p_B)\,D_{AB}$, D_{AB}^2 and $p_A(1 - p_A)\,p_B(1 - p_B)$, respectively, of $\mathbf{x}_{(t)}$. Thus, including recombination,

$$\mathbf{x}_{(t)} = \mathbf{Q}\mathbf{R}\mathbf{x}_{(t-1)}$$

as shown by Hill and Robertson (1968).

We do not propose to list all the transformations possible up to sixth moments, but give some more examples for fourth moments. Consider the case A = D. From Table 2,

$$\mathbf{w}_{(t)} = \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} (1-2p_A)\,\Delta_{ABC} \\ D_{AB}D_{AC} \\ p_A(1-p_A)\,D_{BC} \end{pmatrix}_{(t)}$$

so the transformation $\mathbf{T}$ (Eq. (12)), and generating matrix $\mathbf{Q}$ (Eq. (14)), are again appropriate, but with permutation of rows and columns. However the diagonal elements of the recombination matrix $\mathbf{R}$ are now $[abc]$, $[ab][ac]$ and $[bc]$. Further, let A = C = D, giving

$$\mathbf{x}'_{(t)} = (D_{AB}\,,\, p_A(1 - p_A)\,D_{AB})_{(t)}$$

and A = B = C = D giving

$$\mathbf{x}'_{(t)} = (p_A(1 - p_A),\, p_A{}^2(1 - p_A)^2)_{(t)}\,.$$

Computation of changes in the derived moments is less straightforward with four or more loci. Consider for example, the case of four loci, with A = C and B = D. We have from (3) and Table 2

$$\mathbf{w}_{(t)} = \begin{pmatrix} \delta_{AABB} \\ D_{AB}D_{AB} \\ D_{AA}D_{BB} \\ D_{AB}D_{AB} \end{pmatrix}_{(t)} = \begin{pmatrix} (1-2p_A)(1-2p_B)\,D_{AB} - 2D_{AB}^2 \\ D_{AB}^2 \\ p_A(1-p_A)\,p_B(1-p_B) \\ D_{AB}^2 \end{pmatrix}_{(t)} . \quad (10)$$

If we define the vector $\mathbf{x}_{(t)}$, where

$$\mathbf{x}'_{(t)} = ((1-2p_A)(1-2p_B)\,D_{AB}\,,\, D_{AB}^2\,,\, p_A(1-p_A)\,p_B(1-p_B))_{(t)}\,, \quad (11)$$

we wish to find a recurrence relation for $\mathbf{x}_{(t)}$ in terms of $\mathbf{x}_{(t-1)}$ that does not involve $\mathbf{w}_{(t)}$ and that will enable us to evaluate moments such as D_{AB}^2 each generation. From (10) and (11) we have

$$\mathbf{w}_{(t)} = \mathbf{T}\mathbf{x}_{(t)}\,, \quad (12)$$

where

$$\mathbf{T} = \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

We note that $\mathbf{T}$ is not of full rank, but of full column rank, so it has left inverses $\mathbf{T}^*$, such that

$$\mathbf{T}^*\mathbf{T} = \mathbf{I},$$

with $\mathbf{I}$ being the identity matrix of dimension 3 (Rao and Mitra, 1971). For example, a left inverse of $\mathbf{T}$ is

$$\mathbf{T}^* = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Another which is easy to compute is $\mathbf{T}^* = (\mathbf{T}'\mathbf{T})^{-1}\,\mathbf{T}'$.

Consider firstly just the sampling of chromosome frequencies due to finite population, such that (5) reduces to

$$\mathbf{w}_{(t)} = \mathbf{M}\mathbf{w}_{(t-1)}\,.$$

Using (12) on both sides of the equation and noting that the rank is reduced, we have

$$\mathbf{T}\mathbf{x}_{(t)} = \mathbf{M}\mathbf{T}\mathbf{x}_{(t-1)}\,.$$

A full list of reduction formulae, such as (7), for up to six loci are given in Table 2; in this, the choice of some expressions is arbitrary; for example $(1 - 2p_A)^2$ has been expressed as $1 - 4p_A(1 - p_A)$.

TABLE 2

Equivalence Relations for Reducing Disequilibria

$$D_{AA} = p_A(1 - p_A)$$

$$\Delta_{AAB} = (1 - 2p_A)\, D_{AB}\,,\ \Delta_{AAA} = p_A(1 - p_A)(1 - 2p_A)$$

$$\delta_{AABC} = (1 - 2p_A)\,\Delta_{ABC} - 2D_{AB}D_{AC}\,,\ \delta_{AABB} = (1 - 2p_A)(1 - 2p_B)\, D_{AB} - 2D_{AB}^2$$

$$\delta_{AAAB} = D_{AB} - 6p_A(1 - p_A)\, D_{AB}\,,\ \delta_{AAAA} = p_A(1 - p_A) - 6p_A^2(1 - p_A)^2$$

$$\partial_{AABCD} = (1 - 2p_A)\,\delta_{ABCD} - 2\Sigma_3\, D_{AB}\Delta_{ACD}{}^a$$

$$\partial_{AABBC} = (1 - 2p_A)(1 - 2p_B)\,\Delta_{ABC} - 2\Sigma_2\,(1 - 2p_A)\, D_{AB}D_{BC} - 4D_{AB}\Delta_{ABC}$$

$$\partial_{AAABC} = \Delta_{ABC} - 6p_A(1 - p_A)\,\Delta_{ABC} - 6(1 - 2p_A)\, D_{AB}D_{AC}$$

$$\partial_{AAABB} = (1 - 2p_B)\, D_{AB} - 6p_A(1 - p_A)(1 - 2p_B)\, D_{AB} - 6(1 - 2p_A)\, D_{AB}^2$$

$$\partial_{AAAAB} = (1 - 2p_A)\, D_{AB} - 12p_A(1 - p_A)(1 - 2p_A)\, D_{AB}$$

$$\partial_{AAAAA} = p_A(1 - p_A)(1 - 2p_A) - 12p_A^2(1 - p_A)^2(1 - 2p_A)$$

$$\nabla_{AABCDE} = (1 - 2p_A)\,\partial_{ABCDE} - 2\Sigma_4\, D_{AB}\delta_{ACDE} - 2\Sigma_3\,\Delta_{ABC}\Delta_{ADE}$$

$$\begin{aligned}\nabla_{AABBCD} = {} & (1 - 2p_A)(1 - 2p_B)\,\delta_{ABCD} - 4D_{AB}\delta_{ABCD} - 2\Sigma_2\,(1 - 2p_A)\, D_{AB}\Delta_{BCD} \\ & - 2\Sigma_4\,(1 - 2p_A)\, D_{BC}\Delta_{ABD} + 4\Sigma_2\, D_{AB}D_{AC}D_{BD} - 4\Delta_{ABC}\Delta_{ABD}\end{aligned}$$

$$\begin{aligned}\nabla_{AABBCC} = {} & (1 - 2p_A)(1 - 2p_B)(1 - 2p_C)\,\Delta_{ABC} - 4\Sigma_3\,(1 - 2p_A)\, D_{BC}\Delta_{ABC} \\ & - 2\Sigma_3\,(1 - 2p_A)(1 - 2p_B)\, D_{AC}D_{BC} + 16D_{AB}D_{AC}D_{BC} - 4\Delta_{ABC}^2\end{aligned}$$

$$\nabla_{AAABCD} = \delta_{ABCD} - 6p_A(1 - p_A)\,\delta_{ABCD} - 6\Sigma_3\,(1 - 2p_A)\, D_{AB}\Delta_{ACD} + 12D_{AB}D_{AC}D_{AD}$$

$$\begin{aligned}\nabla_{AAABBC} = {} & (1 - 2p_B)\,\Delta_{ABC} - 6p_A(1 - p_A)(1 - 2p_B)\,\Delta_{ABC} \\ & - 6(1 - 2p_A)(1 - 2p_B)\, D_{AB}D_{AC} - 12(1 - 2p_A)\, D_{AB}\Delta_{ABC} \\ & + 12D_{AB}^2D_{AC} - 2D_{AB}D_{BC} + 12p_A(1 - p_A)\, D_{AB}D_{BC}\end{aligned}$$

$$\begin{aligned}\nabla_{AAABBB} = {} & D_{AB} - 6\Sigma_2\, p_A(1 - p_A)\, D_{AB} + 36p_A(1 - p_A)\, p_B(1 - p_B)\, D_{AB} \\ & + 12D_{AB}^3 - 18(1 - 2p_A)(1 - 2p_B)\, D_{AB}^2\end{aligned}$$

$$\begin{aligned}\nabla_{AAAABC} = {} & (1 - 2p_A)\,\Delta_{ABC} - 12p_A(1 - p_A)(1 - 2p_A)\,\Delta_{ABC} - 14D_{AB}D_{AC} \\ & + 72p_A(1 - p_A)\, D_{AB}D_{AC}\end{aligned}$$

$$\begin{aligned}\nabla_{AAAABB} = {} & (1 - 2p_A)(1 - 2p_B)\, D_{AB} - 12p_A(1 - p_A)(1 - 2p_A)(1 - 2p_B)\, D_{AB} \\ & - 14D_{AB}^2 + 72p_A(1 - p_A)\, D_{AB}^2\end{aligned}$$

$$\nabla_{AAAAAB} = D_{AB} - 30p_A(1 - p_A)\, D_{AB} + 120p_A^2(1 - p_A)^2\, D_{AB}$$

$$\nabla_{AAAAAA} = p_A(1 - p_A) - 30p_A^2(1 - p_A)^2 + 120p_A^3(1 - p_A)^3$$

[a] Sum over possible combinations, e.g. $\Sigma_3\, D_{AB}\Delta_{ACD} = D_{AB}\Delta_{ACD} + D_{AC}\Delta_{ABD} + D_{AD}\Delta_{ABC}$.

b. *Transformations for Higher Moments of Disequilibria*

We are now interested in computing changes in expected values of moments such as D_{AB}^2, Δ_{ABC}^2, $D_{AB}D_{AC}$, etc. It would be possible to derive appropriate moment-generating matrices from first principles, but a shortcut procedure can be based on the above results. Consider, for example, the three-locus case, and let us assume that the alleles A and C are completely associated initially (i.e., are in complete coupling such that if A appears in any chromosome so does C and vice versa) and that there is complete linkage between the loci carrying A and C. We write this as A = C. These alleles remain completely associated, and we can write, at any generation

$$r_{ABC} = q_{AB} = q_{BC},$$
$$q_{AC} = p_A = p_C.$$

The alleles, A and C, are effectively identical, so we can rewrite

$$\Delta_{ABC} = r_{ABC} - p_A q_{BC} - p_B q_{AC} - p_C q_{AB} + 2p_A p_B p_C$$

from Table 1, as

$$\Delta_{ABA} = q_{AB} - p_A q_{BA} - p_B p_A - p_A q_{AB} + 2p_A p_B p_A. \tag{6}$$

Because the ordering of subscripts is immaterial, (6) reduces to

$$\begin{aligned}\Delta_{AAB} &= (1 - 2p_A)(q_{AB} - p_A p_B) \\ &= (1 - 2p_A) D_{AB}\end{aligned} \tag{7}$$

from the definition of D_{AB} (Table 1). If we make the further assumption that A and B are completely associated (i.e., A = B = C),

$$D_{AA} = p_A(1 - p_A) \quad \text{and} \quad \Delta_{AAA} = p_A(1 - p_A)(1 - 2p_A).$$

With complete linkage between A and C, $[abc] = [ab]$. As we are merely considering special cases of starting frequencies and recombinations, the general result (2) also applies in these reduced situations. Therefore

$$\{(1 - 2p_A) D_{AB}\}_{(t)} = (1 - 1/n)(1 - 2/n)[ab]\{(1 - 2p_A) D_{AB}\}_{(t-1)} \tag{8}$$

and

$$\{p_A(1 - p_A)(1 - 2p_A)\}_{(t)} = (1 - 1/n)(1 - 2/n)\{p_A(1 - p_A)(1 - 2p_A)\}_{(t-1)}. \tag{9}$$

Equation (9) can also be derived from the moment-generating formulae for single loci given by Robertson (1952).

TABLE 1

Definitions of Disequilibria

Loci	Chromosomal frequency	Probability of nonrecombinants	
1	p_A		
2	q_{AB}	$[ab]$	$D_{AB} = q_{AB} - p_A p_B$
3	r_{ABC}	$[abc]$	$\Delta_{ABC} = r_{ABC} - \Sigma_3 p_A q_{BC}{}^a + 2p_A p_B p_C$
4	s_{ABCD}	$[abcd]$	$\delta_{ABCD} = s_{ABCD} - \Sigma_4 p_A r_{BCD} - \Sigma_3 q_{AB} q_{CD} + 2\Sigma_6 p_A p_B q_{CD} - 6p_A p_B p_C p_D$
5	t_{ABCDE}	$[abcde]$	$\partial_{ABCDE} = t_{ABCDE} - \Sigma_5 p_A s_{BCDE} - \Sigma_{10} q_{AB} r_{CDE} + 2\Sigma_{10} p_A p_B r_{CDE} + 2\Sigma_{15} p_A q_{BC} q_{DE} - 6\Sigma_{10} p_A p_B p_C q_{DE} + 24 p_A p_B p_C p_D p_E$
6	u_{ABCDEF}	$[abcdef]$	$\nabla_{ABCDEF} = u_{ABDCEF} - \Sigma_6 p_A t_{BCDEF} - \Sigma_{15} q_{AB} s_{CDEF} - \Sigma_{10} r_{ABC} r_{DEF} + 2\Sigma_{15} p_A p_B s_{CDEF} + 2\Sigma_{60} p_A q_{BC} r_{DEF} + 2\Sigma_{15} q_{AB} q_{CD} q_{EF} - 6\Sigma_{20} p_A p_B p_C r_{DEF} - 6\Sigma_{45} p_A p_B q_{CD} q_{EF} + 24\Sigma_{15} p_A p_B p_C p_D q_{EF} - 120 p_A p_B p_C p_D p_E p_F$

[a] Sum over possible combinations, e.g. $\Sigma_3 p_A q_{BC} = p_A q_{BC} + p_B q_{AC} + p_C q_{AB}$.

2. Methods

a. *Definitions and Recurrence Formulae for Mean Disequilibria*

The definitions of gene and gamete frequencies and of the multilocus disequilibria given by Bennett (1954) and used by Hill (1974) are summarised in Table 1, where the letters A, B, etc., refer to a specific allele at loci A, B, etc. These disequilibria are defined such that they decline exponentially in an infinite population, at a rate proportional to the probability of no crossovers between any of the loci concerned. For example with three loci in an infinite population,

$$\Delta_{\mathrm{ABC}(t)} = [abc]\,\Delta_{\mathrm{ABC}(t-1)}\,, \tag{1}$$

where $\Delta_{\mathrm{ABC}(t)}$ is the three-locus disequilibrium and $[abc]$ is the probability that no crossovers occur between the loci A, B and C. As in the previous paper, a haploid model is assumed throughout, with a population comprising n chromosomes. In a finite population, we analyse expected values or moments of the appropriate distribution and find that, for example,

$$E(\Delta_{\mathrm{ABC}(t)}) = (1 - 1/n)(1 - 2/n)[abc]\,E(\Delta_{\mathrm{ABC}(t-1)}) \tag{2}$$

(Hill, 1974). To simplify the succeeding formulae, the *expectations are implicit*, so we rewrite (2)

$$\Delta_{\mathrm{ABC}(t)} = (1 - 1/n)(1 - 2/n)[abc]\,\Delta_{\mathrm{ABC}(t-1)}\,.$$

With four or more loci, simple relations such as (2) are not obtained. For example, with four loci, it is necessary to define a vector $\mathbf{w}_{(t)}$ of expected values at generation t, with transpose

$$\mathbf{w}'_{(t)} = (\delta_{\mathrm{ABCD}}\,, D_{\mathrm{AB}}D_{\mathrm{CD}}\,, D_{\mathrm{AC}}D_{\mathrm{BD}}\,, D_{\mathrm{AD}}D_{\mathrm{BC}})_{(t)}\,. \tag{3}$$

Changes in $\mathbf{w}_{(t)}$ are specified by using the moment-generating matrices $\mathbf{M}$ for drift and $\mathbf{R}$ for recombination, where

$$\mathbf{M} = \frac{n-1}{n^3}\begin{pmatrix} n^2 - 6n + 6 & -2n & -2n & -2n \\ n-1 & n^2 - n & n & n \\ n-1 & n & n^2 - n & n \\ n-1 & n & n & n^2 - n \end{pmatrix} \tag{4}$$

and $\mathbf{R}$ is diagonal, with diagonal elements $[abcd]$, $[ab][cd]$, $[ac][bd]$ and $[ad][bc]$; and

$$\mathbf{w}_{(t)} = \mathbf{MR}\mathbf{w}_{(t-1)}\,. \tag{5}$$

Similar vectors and matrices can be defined for five and six loci (Hill, 1974).

Reprinted from THEORETICAL POPULATION BIOLOGY
All Rights Reserved by Academic Press, New York and London

Vol. 6, No. 2, October 1974
Printed in Belgium

Disequilibrium among Several Linked Neutral Genes in Finite Population

II. Variances and Covariances of Disequilibria

WILLIAM G. HILL

*Statistical Laboratory, Iowa State University, Ames, Iowa 50010**
and
Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN, Scotland†

Received October 1, 1973

A method is derived for computing the variances and covariances of linkage disequilibria between neutral genes in finite populations, which is based on a linear transformation of results given previously for the mean values of disequilibria. The formulae obtained are limited to moments of sixth order or less, such as the variance of the three-locus disequilibrium. It is shown that there is no covariance between any pair of disequilibria in populations starting equilibrium. The pattern of change with time in variance of the three-locus disequilibrium from populations initially in equilibrium is similar to that for two loci, except that the highest values are achieved rather earlier and are smaller.

1. INTRODUCTION

In a previous paper (Hill, 1974), methods were developed for computing expected changes in disequilibria among several linked neutral loci in finite population. Asymptotically, the mean disequilibria approach zero. However, it has already been shown that with pairs of loci having neutral genes there may be a large variance of the disequilibrium so that any sampled population may exhibit considerable disequilibrium (Hill and Robertson, 1968; Sved, 1968; Ohta and Kimura, 1969; and see Kimura and Ohta, 1971, for a review). In this paper formulae for obtaining the variances and covariances of disequilibria among more than two neutral loci are derived.

* Journal paper No. J-7647, Iowa Agriculture and Home Ecomonics Experiment Station, Ames, Iowa, Project 1669. Supported in part by National Institutes of Health, Grant No. 13827.

† Permanent address.

Non-random association of neutral linked genes in finite populations

by

William G. Hill

Reprinted from:
POPULATION GENETICS AND ECOLOGY
© 1976
ACADEMIC PRESS, INC.
New York San Francisco London

Non-Random Association of Neutral Linked Genes in Finite Populations

W.G. HILL

INTRODUCTION

With the widespread use of gel electrophoresis methods to estimate frequencies of polymorphic loci in natural or laboratory populations, information is now also being obtained on linkage disequilibrium between such loci. Using data on disequilibria (an expression of association of gene frequencies at different loci) some additional understanding may be obtained as to the nature of the forces of selection, mutation, drift and migration which help to maintain or reduce polymorphism. Lewontin (1974) has recently reviewed the relevant experimental work and developed these arguments further.

Theoretical and, as yet, experimental studies of linkage disequilibria have mostly been restricted to only pairs of loci. Predictions for approach to equilibria for neutral genes in infinite populations were given by Geiringer (1944) and Bennett (1954). These results have been extended to give expected values of means, variances and covariances of disequilibria for neutral genes in finite populations, but are limited to sixth moments, e.g., the mean disequilibrium among six loci or the variance of disequilibrium at three loci (Hill, 1974a,b). These expectations are computed over all

populations and include those in which one or more loci have reached fixation. It is probably more important to consider disequilibria solely among populations which are segregating at all the relevant loci, and predictions of variation in disequilibria for neutral genes form the subject of this paper. Whilst it was possible to develop analytical methods for disequilibria among all populations, we now have to resort largely to Monte Carlo methods. This parallels previous two locus studies, although some approximations for two-locus disequilibria in segregating populations have been obtained analytically (Sved, 1971; Sved and Feldman, 1973; and see also the review by Kimura and Ohta, 1971).

Effects of selection are not included, the intention being to provide a basis against which selection effects can be compared and also to discuss ways in which data from populations might be analysed. Problems of selection in infinite populations have been reviewed recently for two loci by Karlin (1975) and for more loci have been discussed by Lewontin (1964a,b), Slatkin (1972); Strobeck (1973) and Feldman, Franklin and Thomson (1974). Franklin and Lewontin (1970) also considered selection effects although by simulation in a finite population.

Throughout we shall assume there are just two alleles at each locus; most of the detailed analysis is restricted to three loci, but the extension to more is illustrated and introduces no conceptual difficulties. The extension from two to three loci does introduce some problems, however.

MODEL

Consider loci A,B,C,D,... having that order on a chromosome, and let, for example, a and a' be alternative

alleles at the A locus. There is no mutation and all alleles are neutral with respect to fitness. Gene and chromosome frequencies are denoted p, for example p_a is the frequency of the allele a and $p_{abc'}$ the frequency of the chromosome having alleles a,b and c' . The recombination frequency between loci A and B, for example, is y_{AB}, and there is assumed to be no interference. The map length between these loci is ℓ_{AB}, and since map lengths are additive,

$$\ell_{AC} = \ell_{AB} + \ell_{BC} ,$$

for example. Some comparisons of map length and recombination fraction are shown in Table 1.

TABLE 1. Comparison of $E(r^2_{AB})$ obtained by transition probability matrix iteration for N=8 with that predicted by Sved and Feldman. The ratio of expected values of the moments D^2_{AB} and $p_a p_{a'} p_b p_{b'}$ is also given, as is z_{AB}, (the likelihood ratio statistic x1/2N).

ℓ_{AB}	1/4	1/8	1/16	1/32	1/64	1/128	0
L_{AB}	2	1	0.5	0.25	0.125	0.0625	0
y_{AB}	0.1967	0.1106	0.05875	0.03029	0.01538	0.007752	0
Ny_{AB}	1.5736	0.8848	0.47000	0.24232	0.12304	0.062016	0
$1/(4Ny_{AB}+1)$	0.1371	0.2203	0.3472	0.5078	0.6702	0.8012	1
$1/[1+(4N-2)y_{AB} -(2N-1)y^2_{AB}]$	0.1582	0.2419	0.3689	0.5277	0.6859	0.8119	1
$E(r^2_{AB})$	0.1614	0.2620	0.4328	0.6322	0.7898	0.8878	1
$\frac{E(D^2_{AB})}{E[p_a p_{a'} p_b p_{b'}]}$	0.1776	0.2880	0.4638	0.6578	0.8062	0.8970	1
z_{AB}	0.1810	0.2882	0.4688	0.6805	0.8493	0.9549	1.0766

The population comprises N diploid individuals, and products of population size and map length are denoted L, e.g., $L_{AB} = N\ell_{AB}$. Generations, t, are non-overlapping.

A haploid model is used to reproduce the populations. From the 2N chromosomes of one generation the frequency distribution is squared to give the expected genotypic frequencies after random mating including random selfing. The expected gametic output after recombination is computed, and the next generation obtained by sampling 2N chromosomes from the multinomial distribution. This procedure is used exactly in the algebraic analysis (Hill, 1974a), and in the Monte Carlo method it is simulated by sampling 2N pairs of parental chromosomes with replacement, and sampling from each pair a recombinant progeny chromosome.

REVIEW OF TWO-LOCUS THEORY

Some difficulties are encountered in defining and testing for association of gene frequency at three or more loci. Thus we consider some of the alternatives and initially review the two locus theory.

If there is dependence of frequencies at the two loci, $p_{ab} \neq p_a p_b$, and the measure of disequilibrium commonly used is

$$D_{AB} = p_{ab} - p_a p_b = p_{ab}p_{a'b'} - p_{ab'}p_{a'b} \quad . \qquad (1)$$

For neutral genes in infinite populations at generation t

$$D_{AB(t)} = (1-y_{AB})D_{AB(t-1)}$$

and for finite populations, taking expectations over a conceptual set of identical populations,

$$E(D_{AB(t)}) = (1-1/2N)(1-y_{AB})D_{AB(t-1)} \qquad (2)$$

for the haploid model (Wright, 1933). Thus the mean disequilibrium always approaches zero.

The disequilibrium can also be viewed as the covariance of gene frequencies. For example, Slatkin (1972) expressed it as

$$D_{AB} = E[(x_a - p_a)(x_b - p_b)] \qquad (3)$$

where x_a and x_b are the number (i.e., 0 or 1) of a and b genes, respectively, on the chromosome, and expression (3) is equivalent to (1). Based on the covariance concept, a useful measure of two locus disequilibrium is the correlation, r_{AB}, or squared correlation of gene frequencies

$$r^2_{AB} = D^2_{AB}/(p_a p_{a'} p_b p_{b'}) \quad , \qquad (4)$$

which is defined only for segregating populations (Hill and Robertson, 1968). The range of possible values of r^2_{AB} is much less dependent on gene frequencies than is D_{AB}, although, as Sved (1971) has pointed out, r^2_{AB} cannot reach values of unity for many combinations of gene frequencies. The property of r^2_{AB} of which we shall make most use in this paper is its relation to the chi-square statistic in the contingency table test for association between alleles at the A and B locus when $2N$ chromosome types are identified. This statistic is $2Nr^2_{AB}$, so in the first generation of finite population started from a population in equilibrium, $2Nr^2_{AB}$, measured among the sampled parental chromosomes, has an approximately χ^2 distribution with 1 d.f., and thus r^2_{AB} has a mean of $1/2N$ (Hill and Robertson, 1968). In a biological population the test for association has usually to be carried out from a sample of progeny whose numbers differ from that of the parents. Also, as a rule, the data available are on diploids so that chromosome frequencies have to be estimated by maximum likelihood. However, for codominant loci, the statistic for testing for association is nr^2_{AB}, where n are the number of chromosomes (in the haploid case) or diploid

individuals sampled (Hill, 1974c).

As shown by (2), the mean disequilibrium over populations approaches zero for neutral genes, but as a result of sampling there is a variance in D_{AB} between populations. The asymptotic value of $E(r^2_{AB})$ taken over segregating populations was shown by Hill and Robertson (1968) to equal $1/4Ny_{AB}$ approximately, for large y_{AB}, and to approach unity for $y_{AB} = 0$. Subsequently, Sved (1971) found values for $E(r^2_{AB})$ using an argument which is rather hard to follow, and his result was modified by Sved and Feldman (1973) to give

$$E(r^2_{AB}) = 1/[1+(4N-2)y_{AB} - (2N-1)y^2_{AB}] \quad , \qquad (5)$$

which simplifies to

$$E(r^2_{AB}) \sim 1/(1+4Ny_{AB}) \qquad (6)$$

for large N and small y_{AB}. It is clear that (6) is correct either when Ny_{AB} is very large or approaches zero, but we have undertaken some numerical checks for other values. For populations with $N \leq 8$ this was done by transition probability matrix iteration, with the matrix kept to small size by utilising the symmetry of the model in a program described elsewhere (Hill, 1969); for larger values of N Monte Carlo simulation was used. Some typical exact results are given in Table 1 for $N=8$. It is clear that while the Sved-Feldman formula (5) gives a good general impression of $E(r^2_{AB})$, it is not formally correct. The largest differences occur around $Ny_{AB} = 0.25$, the Sved-Feldman values being underestimated by some 20%. The table also shows that even for this small N value, the more involved formula (5) is little better than the approximation (6). Similar results were obtained using $N=4$ and $N=6$, by matrix iteration. Simulation for $N\ell_{AB} = 0.25$ and 0.5, corresponding to $Ny_{AB} = 0.246$ and 0.488, respectively, was carried out with $N=20$ and 6400 replicates,

giving values of $E(r^2_{AB})$ of about 0.62 and 0.39, compared with the Sved-Feldman predictions from (5) of 0.51 and 0.35, respectively. It seems clear that, with increasing population size, the value of $E(r^2_{AB})$ will not asymptote at $1/(1+4Ny_{AB})$. The prediction of changes with generation in $E(r^2_{AB})$ in populations starting at equilibrium, which can be obtained from Sved and Feldman (1973), are also somewhat in error. (We are not sure where the logic of the Sved-Feldman approach breaks down, one possibility is in the assumption that the same probabilities of identity at the two loci apply in segregating and non-segregating populations.)

The moments $E(D^2_{AB})$ and $E(p_a p_{a'} p_b p_{b'})$ over all populations, whether segregating or not, are more readily computed than $E(r^2_{AB})$, either by iteration of a moment generating matrix (Hill and Robertson, 1968) or by a diffusion approximation (Ohta and Kimura, 1969). As Ohta and Kimura (1969) showed by Monte Carlo simulation, however, the steady state value of the ratio $\sigma^2_d = E(D^2_{AB})/E(p_a p_{a'} p_b p_{b'})$ over all populations is a very good approximation to the steady state value of $E(r^2_{AB})$, computed only in segregating populations. A further illustration, using exact values from the transition probability matrix for $N=8$, of the similarity of these two quantities is given in Table 1. Also the approaches to the steady state values of $E(r^2_{AB})$ and σ^2_d with increasing generations are very similar.

Whilst not of particular relevance to the two locus review, we introduce a further measure of disequilibrium which will be useful for more loci. An alternative to the chi-square method of testing for association in a contingency table and thus for gene linkage disequilibrium is the likelihood ratio test (e.g., Sokal and Rohlf, 1969). The test statistic for chi-square may be written

$$\Sigma(\text{observed} - \text{expected})^2/\text{expected},$$

whereas that for the likelihood ratio is

$$2\,\Sigma(\text{observed})\,\log_e(\text{observed/expected}). \qquad (7)$$

In this context, the likelihood ratio statistic (7) can be written in terms of population size and frequencies as

$$2Nz_{AB} = 4N[p_{ab}\log(p_{ab}/p_ap_b) + \ldots + p_{a'b'}\log(p_{a'b'}/p_{a'}p_{b'})]$$

giving

$$z_{AB} = 2[p_{ab}\log p_{ab} + \ldots + p_{a'b'}\log p_{a'b'} - p_a\log p_a - \ldots - p_b\log p_{b'}] \qquad (8)$$

(where we are dealing with $2N$ identified chromosomes). In a sample of parents from a population in linkage equilibrium, the likelihood ratio $2Nz_{AB}$ is asymptotically (for large N) distributed as χ^2 with 1 d.f. and it is easy to show that

$$z_{AB} = r_{AB}^2 + \text{terms in } D^3, D^4, \ldots$$

It turns out, however, that z_{AB} and r_{AB}^2 are numerically very similar over a wide range of parameters, and $E(z_{AB}^2)$ taken over segregating populations is closely approximated by $E(r_{AB}^2)$, as demonstrated in Table 1.

MEASURES OF DISEQUILIBRIUM FOR THREE OR MORE LOCI

A three locus disequilibrium, D_{ABC}, was defined by Bennett (1954) and equals that of Slatkin (1972). Extending (3),

$$D_{ABC} = E[(x_a - p_a)(x_b - p_b)(x_c - p_c)]$$

$$= p_{abc} - p_ap_{bc} - p_bp_{ac} - p_cp_{ab} + 2p_ap_bp_c \quad . \qquad (9)$$

It has been shown that

$$E(D_{ABC(t)}) = (1-1/2N)(1-1/N)(1-y_{AB})(1-y_{BC})D_{ABC(t-1)} \qquad (10)$$

(Hill, 1974a), and, of course, (10) gives the infinite population result of Bennett as $N \to \infty$.

A test for association of gene frequencies now involves a

2^3 contingency table and assuming that the gene frequencies, which are the marginal frequencies of the table, are estimated in the same analysis there are a total of 4 degrees of freedom available for testing the hypothesis of independence, $p_{abc} = p_a p_b p_c$. A partition of these 4 d.f. was suggested (in a general rather than genetic context) by Lancaster (1951) and follows the usual analysis of variance of a 2^3 factorial: use 1 d.f. each for testing pairs of loci together, i.e., the three hypotheses $p_{ab} = p_a p_b$, $p_{ac} = p_a p_c$ and $p_{bc} = p_b p_c$ in each case summing over frequencies at the third locus, and attribute the residual chi-square to the 1 d.f. for three-locus association or disequilibrium. If we denote by X^2 the chi-square statistic with 4 d.f. for testing $p_{abc} = p_a p_b p_c$, which is

$$X^2 = 2N[(p_{abc}-p_a p_b p_c)^2/(p_a p_b p_c) + \ldots + (p_{a'b'c'}-p_{a'}p_{b'}p_{c'})^2/(p_{a'}p_{b'}p_{c'})] \quad , \qquad (11)$$

it can be shown that

$$X^2 = 2N[r^2_{AB} + r^2_{AC} + r^2_{BC} + r^2_{ABC}] \quad , \qquad (12)$$

where, by analogy with (4),

$$r^2_{ABC} = D^2_{ABC}/(p_a p_a' p_b p_b' p_c p_c') \qquad (13)$$

and D_{ABC} is given by (9). The quantity r^2_{ABC} is not a correlation as such, but a natural extension of the correlation concept to three variables. Thus it appears at first sight that Lancaster's partition can be interpreted immediately in terms of two-locus disequilibria; and in a sample of $2N$ chromosomes from a population in equilibrium $2Nr^2_{ABC}$ is approximately distributed as χ^2 with 1 d.f.

The partition due to Lancaster has been criticised on several grounds, however, initially by Plackett (1962) who

argued that Lancaster's criterion was not strictly a test of three-way (three-locus) association. He considered the necessary criterion to be that the association between, say, A and B should be the same in the group (chromosomes) having c as c' , and this criterion should be the same if we consider A with C and B with C in addition. The criterion proposed by Bartlett (1935) to define no three-way association,

$$p_{abc}p_{ab'c'}p_{a'bc'}p_{a'b'c} = p_{abc'}p_{ab'c}p_{a'bc}p_{a'b'c'} \qquad (14)$$

does satisfy Plackett's conditions: it is symmetric among the loci, and expressing the association between A and B at each level of C , (14) gives

$$\frac{p_{abc}p_{a'b'c}}{p_{ab'c}p_{a'bc}} = \frac{p_{abc'}p_{a'b'c'}}{p_{ab'c'}p_{a'bc'}} \quad .$$

An example given by Plackett (1962) illustrates that Bartlett's criterion (14) for no three-way association is not the same as that due to Lancaster, which in genetical terms is equivalent to $D_{ABC} = 0$ (from (9)). If the frequencies, each multiplied by 24, are

$$p_{abc} = 1 \ , \ p_{abc'} = 2 \ , \ p_{ab'c} = 2 \ , \ p_{ab'c'} = 6 \ ,$$

$$p_{a'bc} = 3 \ , \ p_{a'bc'} = 2 \ , \ p_{a'b'c} = 4 \ , \ p_{a'b'c'} = 4 \ ,$$

equation (14) is satisfied, but $D_{ABC} \neq 0$. A more extreme example, but with the same outcome, which we shall find relevant to our subsequent discussion, is $p_{a'b'c'} = 1 - p_{abc}$ with all other frequencies equal to zero.

While the definition of disequilibrium among three loci given by (9) and the partition of chi-square in (12) have some appeal they do not conform with current methods of analysis of three-way tables, which more or less uniformly are based on the Bartlett-Plackett model (see e.g., Goodman (1969) and

Fienberg (1970) for exposition). In the three-way model a hierarchy of levels of independence among the frequencies can be constructed which can help interpretation (Goodman, (1969); Fienberg, (1970)); and in genetical language, a typical hierarchy is shown in Table 2, which follows Hill (1975).

TABLE 2.

a. Succession of models of association of gene frequencies at three loci.

Model	p_{abc}	Fitted Association	Log likelihood[g]
0 : Complete independence of frequencies	$p_a p_b p_c$	-	K(A)+K(B)+K(C)
1 : Frequencies at C independent of A & B†	$p_{ab} p_c$	AB	K(AB)+K(C)
2 : Independence at B & C conditional on A†	$p_{ab} p_{ac} / p_a$	AB,AC	K(AB)+K(AC)-K(A)
3 : No three-way association	p^*_{abc}‡	AB,AC,BC	K*(ABC)
4 : All associations	p_{abc}	AB,AC,BC,ABC	K(ABC)

b. Succession of likelihood ratio statistics, each with 1 d.f.††

Source (difference in models)	Fitted Association	Log Likelihood ratio
Marginal assoc. A & B (1-0)	AB	$2Nz_{AB} = 2[K(AB)-K(A)-K(B)]$
Assoc. A & C given assoc. A & B (2-1)	AC	$2Nz_{AC} = 2[K(AC)-K(A)-K(C)]$
Assoc. B & C given assoc. A & B and A & C (3-2)	BC	$2Nz'_{BC} = 2[K^*(ABC)-K(AB)-K(AC)+K(A)]$
Assoc. A, B & C given pair-wise assoc. (4-3)	ABC	$2Nz'_{ABC} = 2[K(ABC)-K^*(ABC)]$
Total		$2Nz_{ABC} = 2[K(ABC)-K(A)-K(B)-K(C)]$

† One of three, †† One of six, alternative hierarchies

‡ No explicit formula for p^*_{ABC}, but satisfies (14)

g $K(A) = 2N(p_a \log p_a + p_{a'} \log p_{a'})$

$K(AB) = 2N(p_{ab} \log p_{ab} + \ldots + p_{a'b'} \log p_{a'b'})$

$K(ABC) = 2N\Sigma p_{abc} \log p_{abc}$, $K^*(ABC) = 2N\Sigma p^*_{abc} \log p^*_{abc}$ (sum over 8 types)

This shows in Table 2a the values which would be taken by $p_{abc}, \ldots, p_{a'b'c'}$ when there is no association of frequencies, one pair is associated, and so on. There is no explicit formula for chromosome frequencies (p^*_{abc}) when there are all pair-wise but no three-way associations, with (14) satisfied. An iterative routine, however, given by Fienberg (1970) for example, can be used to obtain these values of p^*_{abc}. Also given in Table 2a are the log likelihoods, K, apart from constant terms, obtained by fitting the different models. For model 4 in which all two-way and three-way associations are fitted,

$$2N\Sigma\, p_{abc} \log p_{abc} = K(ABC) \qquad (15)$$

where summation is over all chromosome types. For model 2 in which, for example, B and C are independent, conditional on the gene at A, the expected frequencies satisfy $p_{abc} = p_{ab}p_{ac}/p_a$. Estimates of these pair-wise frequencies are given by marginal totals, e.g., $p_{ab} = p_{abc} + p_{abc'}$, so the log likelihood becomes, from (15),

$$\begin{aligned} 2N\Sigma p_{abc} \log(p_{ab}p_{ac}/p_a) &= \\ &= 2N(\Sigma p_{ab} \log p_{ab} + \Sigma p_{ac} \log p_{ac} - \Sigma p_a \log p_a) \\ &= K(AB) + K(AC) - K(A) \quad , \end{aligned} \qquad (16)$$

say, where K(AB) and K(A) denote log likelihoods computed from the specified marginal totals. For example $K(A) = p_a \log p_a + p_{a'} \log p_{a'}$. Using these likelihoods a succession of models can be fitted as shown in Table 2b, that given being one of 6 alternative sequences. For example, the test statistic $(2Nz'_{BC})$ for association between B and C, assuming association between A and B and between A and C is given by the likelihood ratio, or difference in log likelihoods, from fitting model 3 (all pair-wise but no three-way

associations) versus model 2 (B and C independent, but conditional on the level of A, which implies possible associations between A and B and between A and C). It turns out (see Table 2) that the likelihood ratio test for association between a pair, A and B say, is the same if no other pairs are fitted previously or one pair is fitted previously. This ratio is denoted $2Nz_{AB}$, and is, of course, equal to $2N \times$ that given by equation (8). The likelihood ratio statistic for testing for three-way association is denoted $2Nz'_{ABC}$, and the total with 4 d.f. is denoted $2Nz_{ABC}$. If there is no association, each statistic is asymptotically χ^2 distributed. For further details of interpretation see Smouse (1974) and Hill (1975).

It would be possible to undertake the same partition as shown in Table 2a and analyse by the traditional $[\Sigma \text{ (observed - expected)}^2/\text{expected}]$ chi-square analysis. However we shall use likelihood ratios to quantify this partition since they show more desirable properties in finite populations in which there are considerable departures from equilibrium after a few generations.

Four or more loci

The four locus disequilibrium defined by Bennett (1954) and Hill (1974a) differs from that of Slatkin (1972), whose is more closely related to chi-square. Extending (3) to four loci we obtain

$$D_{ABCD(t)} = p_{abcd} - p_a p_{bcd} - \cdots - p_d p_{abc} + \\ + p_a p_b p_{cd} + \cdots + p_c p_d p_{ab} - 3p_a p_b p_c p_d \quad . \qquad (17)$$

Changes in $D_{ABCD(t)}$ cannot be expressed in the simple form of (2), but expressions for change in the vector $(D_{ABCD}, D_{AB}D_{CD}, D_{AC}D_{BD}, D_{AD}D_{BC})$ can be given (Hill, 1974a).

Defining $r^2_{ABCD} = D^2_{ABCD}/(p_a \ldots p_d,)$, the chi-square statistic extending (12) to four loci can be shown to be

$$2N[r^2_{AB} + \ldots + r^2_{CD} + r^2_{ABC} + \ldots + r^2_{BCD} + r^2_{ABCD}]$$

having 11 terms, each corresponding to 1 d.f.

The likelihood ratio partition, extending that in Table 2, is given by Goodman (1970), but in view of the computational requirement in many replicates of Monte Carlo simulation we shall restrict discussion to the total departure from equilibrium in which all possible associations are fitted. By analogy with Table 2b this quantity is

$$2[K(ABCD)-K(A)-K(B)-K(C)-K(D)] = 2Nz_{ABCD} \quad . \qquad (18)$$

LIMITING PREDICTIONS

Before embarking on detailed simulation results it is useful to consider a few special cases analytically. For two loci, the ratio of expectations of moments $\sigma^2_d = E(D^2_{AB})/E(p_a p_a, p_b p_b,)$ turned out to be a good predictor of r^2_{AB} , the expectation of the ratio of these quantities (Ohta and Kimura, 1969; Table 1 of this paper). Unfortunately the equivalent result does not always apply for three loci, as comparison of results of Hill (1974b) and the simulation to be described will show. Thus our analytical results are very limited, but give some useful insight into the multi-locus problems.

When population size and recombination fractions are sufficiently large that $Ny_{AB} > 1$, approximately, a simple argument was used by Hill and Robertson (1968) to show that $E(r^2_{AB}) = 1/4Ny_{AB}$, approximately, at the steady state. They argued that providing $r^2_{AB} = D^2_{AB}/p_a p_a, p_b p_b,$ was small, it was

reduced to $(1-y_{AB})^2 r^2_{AB} \sim (1-2y_{AB}) r^2_{AB}$ by recombination and increased by $1/2N$ by drift each generation, since $2Nr^2_{AB}$ is asymptotically χ^2 with 1 d.f. and has an expectation of unity in samples from populations in equilibrium. Equating the increase and loss gives $E(r^2_{AB}) = 1/4Ny_{AB} \sim 1/4L_{AB}$, the substitution of map length for recombination fraction being adequate for $y_{AB} < 0.1$ or so (Table 1). This argument can be extended to three loci. Consider $r^2_{ABC} = D^2_{ABC}/p_a p_a, p_b p_b, p_c p_c$, with Ny_{AB} and Ny_{BC} large. The loss due to recombination is by the factor $(1-y_{AB})^2(1-y_{BC})^2 \sim 1 - 2(y_{AB}+y_{BC})$ and, using χ^2, the increment due to drift is $1/2N$, giving $E(r^2_{ABC}) = 1/4N(y_{AB}+y_{BC}) \sim 1/4L_{AC}$. Thus the steady state value of $E(r^2_{ABC})$ is approximately equal to that of $E(r^2_{AC})$, i.e., the two-locus measure of disequilibrium between the outer pair of loci on the chromosome. The value of $1/2N \times$ the total chi-square for disequilibrium at three loci is, using (11) and (12),

$$E(r^2_{AB}+r^2_{BC}+r^2_{AC}+r^2_{ABC}) = (L_{AB}^{-1}+L_{BC}^{-1}+2L_{AC}^{-1})/4 \quad , \qquad (19)$$

which reduces to $3/2L_{AC}$ if $L_{AB} = L_{BC} = \frac{1}{2}L_{AC}$. Similar results can be obtained for four or more loci.

We now turn to the special case of no recombination amongst the loci. The population then comprises a set of different chromosomes which behave like neutral alleles, and, regardless of the number of loci, eventually only two types will remain segregating (Kimura, 1955). In some replicate populations these will comprise segregants at all the loci in question, which for three loci would be the pairs abc/a'b'c', abc'/a'b'c' etc. Quantities such as r^2_{ABC} and likelihood ratio statistics such as $2Nz_{ABC}$ then depend only on the frequency of the alternative chromosomes, and not on their specific configuration. Thus assume abc/a'b'c' are segre-

gating, and let $p_{ABC} = p$. Then, from (9),

$$D_{ABC} = p(1-p)(1-2p) \quad , \tag{20}$$

and from (13),

$$r^2_{ABC} = (1-2p)^2/[p(1-p)]$$
$$= 1/p + 1/(1-p) - 4 \tag{21}$$

which is symmetric about $p = 0.5$. Single locus theory can be used to find $E(r^2_{ABC})$ in (21). Using a transition probability matrix with elements q_{ij} specifying the probability the population has j chromosomes of type abc in generation $t+1$ given that it had i at time t , with

$$q_{ij} = \binom{2N}{j} (i/2N)^j (1-i/2N)^{2N-j} \quad ,$$

asymptotic values of $E(r^2_{ABC})$ were obtained and are given in Table 3.

TABLE 3. Asymptotic values of $E(r^2_{ABC})$, $E(r^2_{ABCD})$ and $E(z_{AB})$ with no recombination computed from the exact distribution or by the approximation using the uniform distribution.

	N	10	20	40	60
$E(r^2_{ABC})$	exact	3.17	4.34	5.58	6.33
	approx.	2.54	3.71	4.96	5.72
$E(r^2_{ABCD})$	exact	40.8	93.4	202.8	314.3
$E(z_{AB})$	exact	1.064	1.037	1.021	1.015
	approx.	1.086	1.046	1.024	1.019

The steady state distribution of unfixed classes is approximately uniform (Fisher, 1930), so a simple solution is

$$E(r^2_{ABC}) = \int_{1/2N}^{1-1/2N} \left(\frac{1}{p} + \frac{1}{1-p} - 4\right) dp \Big/ \int_{1/2N}^{1-1/2N} dp$$

$$= 2[\log 2N + \log(1-1/2N)]/(1-1/N) - 4 , \qquad (22)$$

which approaches $2 \log 2N-4$ as N increases. These results are compared with the exact values in Table 3. The choice of $1/2N$ and $1-1/2N$ for the bounds is somewhat arbitrary, and the approximation (22) can be improved by modifying the bounds and allowing for the slight departure from uniformity found at the ends of the distribution (Fisher, 1930). From (12), and noting that r^2_{AB}, r^2_{AC} and r^2_{BC} equal unity when there are only two chromosome types, the total chi-square expected is given by $2N[E(r^2_{ABC}) + 3]$.

The approximate and exact values of $E(r^2_{ABC})$ shown in Table 3 increase in parallel, as $\log N$. This contrasts with $E(r^2_{AB})$ for two loci which asymptotes at a value of unity for any population size. The same calculations done for four loci with no recombination give from (17) when only two chromosome types are segregating,

$$r^2_{ABCD} = \frac{1}{p^2} + \frac{1}{(1-p)^2} - \frac{4}{p(1-p)} + 9 .$$

Exact values of $E(r^2_{ABCD})$ obtained using the transition matrix are given in Table 3. The approximation, obtained by integrating over the uniform distribution, suggests that $E(r^2_{ABCD})$ increases in proportion to N , a result largely borne out by the exact values. The asymptotic values of the chi-square statistics for two, three and four loci thus increase as N , $N \log N$ and N^2 , respectively, when there is no recombination.

An extension of diffusion equation arguments, such as those of Hill and Robertson (1966) or Ohta and Kimura (1969), for

two loci to three or more, suggest that, on a time scale inversely proportional to N, the distribution of chromosome types is independent of N and a function of the vector of N x recombination fractions or map lengths, providing the recombination fractions are of order $1/N$. Thus the expected value of any quantity, such as r^2_{ABC}, which is a function of frequencies (equation 13) and not of population size, might be expected to be independent of N. However, as (21) shows, in the limiting case of no recombination, r^2_{ABC} has terms in $1/p$ and $1/(1-p)$ and takes its highest values at the ends of the distribution. The value of $1/p$ at the end point in a finite population with discrete classes is proportional to N, and thus it is clear that the continuous diffusion approximation does not hold there.

The likelihood ratio statistic for testing for all departures from random association with two loci is equal to $2Nz_{AB}$ (equation 8) and when only two chromosome types are segregating with frequencies p and $1-p$, z_{AB} reduces to

$$z_{AB} = 2p \log p - 2(1-p) \log (1-p) \qquad (23)$$

from (8), taking $p \log p = 0$ as $p \to 0$. Integrating z_{AB} over the uniform distribution as in (22) and using its symmetry about $p = 0.5$, we obtain

$$E(z_{AB}) = -4 \left(\int_{1/2N}^{1-1/2N} p \log p \, dp \right) / (1-1/N)$$

$$= \left\{ [p^2(1-2 \log p)] \right\}_{1/2N}^{1-1/2N} / (1-1/N) \quad . \qquad (24)$$

The approximation in (24) can be further approximated to $E(z_{AB}) = 1+1/N$, and tends to unity for large N, which is also the asymptotic value of $E(r^2_{AB})$ for two loci.

Now let us consider three loci. The likelihood ratio statistic for testing for all departures from random association is $2Nz_{ABC}$ with 4 d.f. (Table 2) and with only two chromosome types segregating is proportional to

$$z_{ABC} = 2[p \log p^3 + (1-p)\log(1-p)^3 - 3p \log p - 3(1-p)\log(1-p)]$$
$$= -4[p \log p + (1-p) \log(1-p)]$$
$$= 2z_{AB} \qquad (25)$$

from (23), since the expected frequency of a chromosome such as abc is p^3 if a , b and c each have frequency p . Thus, using (24), for large N

$$E(z_{ABC}) = 2 \qquad (26)$$

asymptotically, when there is no recombination, and is not a function of N as are the equivalent quantities of the standard chi-square statistics: whereas z_{ABC} includes terms in p log p which tend to zero as p becomes very small, r^2_{ABC} includes terms in 1/p . The approximation (24) using the uniform distribution is compared with the exact value using the transition matrix in Table 3. The agreement is very good, as is the further approximation 1+1/N .

Using Table 2 we find that the likelihood ratio for the asymptotic case of three loci with no recombination and large population size would be partitioned as follows:

Source	Log likelihood ratio x (1/2N)
Marginal assoc. of A & B	$z_{AB} = 1$
Assoc. of A & C after A & B	$z_{AC} = 1$
Assoc. of B & C after A & B and A & C	$z'_{BC} = 0$
Assoc. A, B and C after all pairs	$z'_{ABC} = 0$
Total	$z_{ABC} = 2$

Of course, any other sequence of fitting the pairs would give the same partition in the sense that the first two pairs fitted would account for the total likelihood ratio.

Continuing to four loci and extending (25), we find that the total likelihood ratio statistic is three times that for two loci, and in general with no recombination the total likelihood ratio statistic is proportional to the number of pairs of "adjacent" loci, i.e., one less than the number of loci.

In view of the more desirable behaviour of the likelihood ratio statistics over chi-square statistics with change in population size at very low recombination values, as illustrated by these results for small population size, most of the remainder of the results will be restricted to them.

SIMULATION RESULTS

All simulations were started with linkage equilibrium, gene frequencies of 0.5 and, unless noted to the contrary, with 1600 replicates, and all results are plotted solely for replicates in which all loci are segregating at the specified generation. It is regrettable that most precise information is obtained on early generations, before many replicates have been fixed, and steady-state values of quantities such as likelihood ratios cannot be obtained accurately without excessive computing expenditure. Procedures which start with populations sampled from one segregating after many generations and thus representative of the steady state, have to be used with caution to avoid introducing new biases. The main advantage in commencing with equilibrium per se is that the results then show the rate at which disequilibrium accumulates by chance.

Three loci: effect of change in population size

Comparisons are given in Figure 1 of total likelihood ratio statistics (expressed as z_{ABC}, i.e., likelihood ratio /2N) for three equally spaced loci in populations of size 10, 20 and 40.

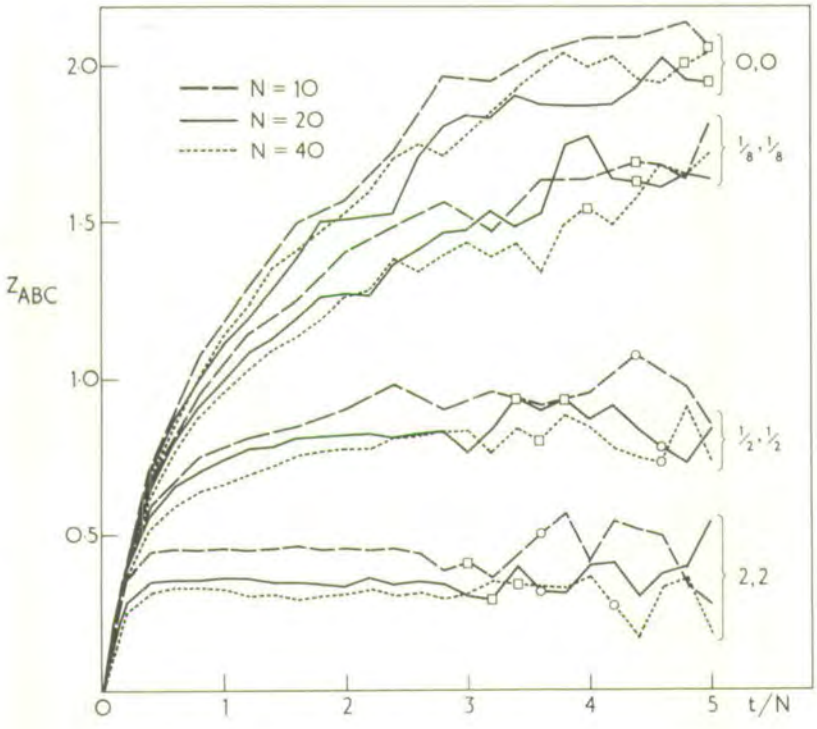

FIGURE 1. *Expected values of z_{ABC} (where z_{ABC} = total likelihood ratio/2N) for three neutral loci, initially in linkage equilibrium with frequency 0.5. Generations (t) are plotted as a proportion of population size (N). Results are given for several values of N×map lengths (L_{AB},L_{BC}) with computations made at three values of N. There are initially 1600 segregating replicates, □ , 0 denotes < 50 , < 20 segregating respectively.*

Generations, on the abscissa, are plotted on a scale proportional to N, so that results for different population sizes and the same values of $L = N \times$ map length can be directly compared. It is seen that, especially at the highest distances apart ($L_{AB} = L_{BC} = 2$) the expected value of z_{ABC} in segregating populations is rather higher at N=40 than N=10. Nevertheless the changes induced in z_{ABC} by four-fold changes in N at constant map length (e.g., N=10 or 40, $\ell_{AB} = 0.05$ giving $L_{AB} = 0.5$ or 2) are much greater than four-fold changes in N at constant $N \times$ map length (e.g., N=10 or 40, $\ell_{AB} = 0.05$ or 0.0125 giving $L_{AB} = 0.5$). Although computations at larger population sizes than 40 were undertaken for the smaller L values because of computing expense, it seems reasonably safe from the theoretical arguments and results of Figure 1 to deduce values for larger population sizes from simulations with very small ones. (This conclusion would not have held if values of r^2_{ABC} had been plotted at low L values.) Subsequent figures in which partitions of z_{ABC} have been made, or in which more than three loci have been included have all been carried out with N=20 for low values of L or N=80 at high values of L where equilibrium is reached relatively earlier. It is noted in Figure 1 that there is closer agreement between simulation results at N=20 and N=40 than between N=20 and N=10 .

Three loci: partition of likelihood ratio

Partitions of z_{ABC} are given in Figure 2 for six different pairs of values of L_{AB} and L_{BC} using N=20 and in Figure 3 for three more pairs using N=80 and less replication (400). When two or more quantities have the same expected value, for example z_{AB} and z_{BC} when $L_{AB} = L_{BC}$, their mean is plotted.

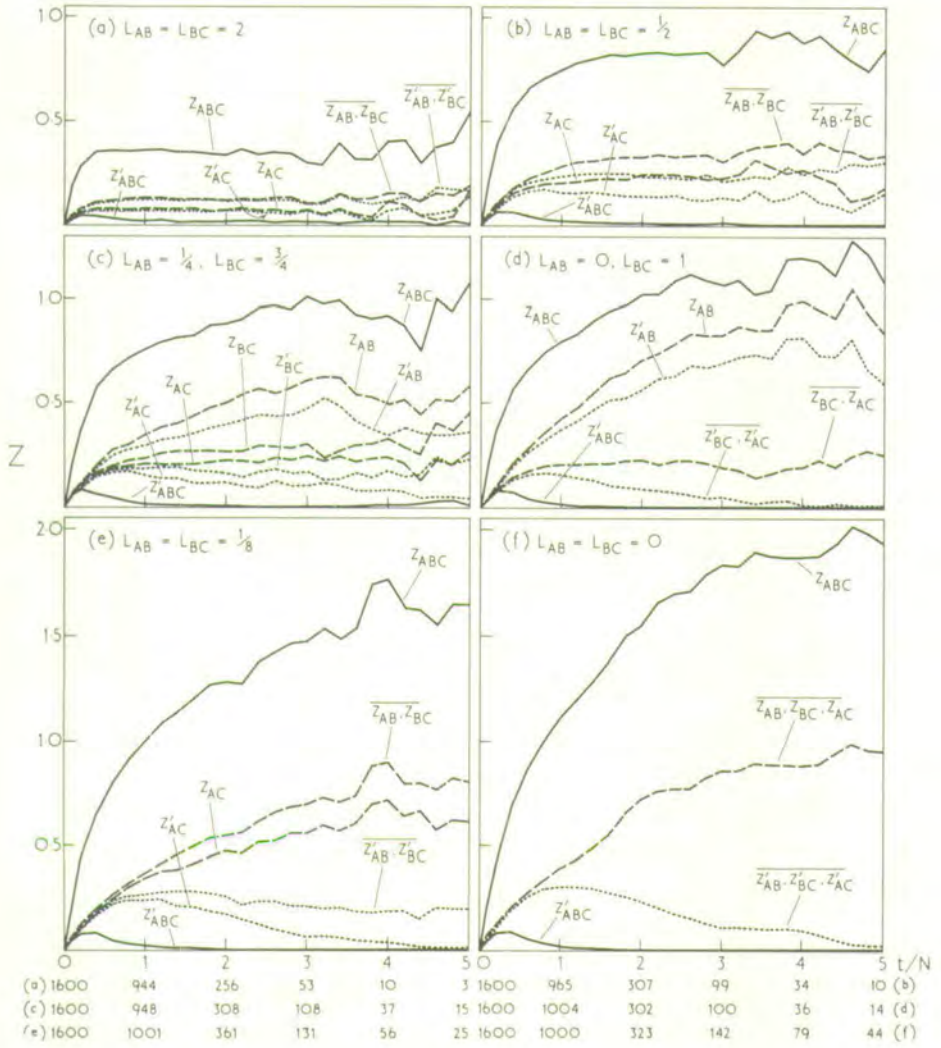

FIGURE 2. Expected values of partitioned values of z (=likelihood ratio/2N) plotted against t/N for three neutral loci initially in linkage equilibrium with frequency 0.5 and population size N=20. Numbers of segregating replicates are shown. Results for different values of N×map lengths (L_{AB},L_{AC}) are a:(2,2), b:(0.5,0.5), c:(0.25,0.25), d:(0,1), e:(0.125,0.125), f:(0,0).

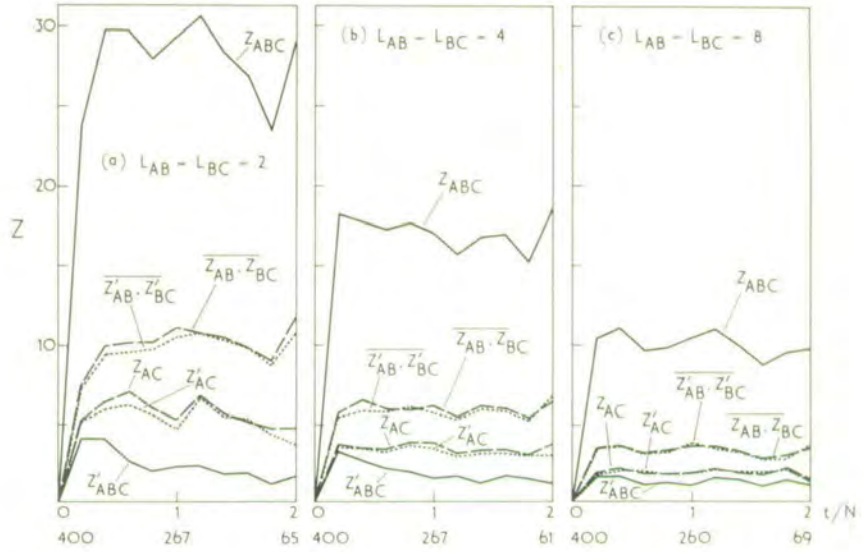

FIGURE 3. *As Figure 2, but with N=80, larger values of N × map lengths and 400 replicates. (L_{AB},L_{BC}) are a:(2,2), b:(4,4), c:(8,8).*

Consider firstly the case of no recombination (L_{AB}=L_{BC}=0). In Figure 2f quantities such as z_{AB} asymptote at approximately 1.0, z_{ABC} at 2.0 and z'_{AB} and z'_{ABC} at 0.0 as predicted by (26) and the discussion following the equation. However about 5N generations (i.e., 100 generations with N=20) are required before these asymptotic values are approached. In the first generation from equilibrium the quantities 2Nz are asymptotically χ^2 distributed, and so z_{ABC}, z_{AB}, z'_{AB} and z'_{ABC} have expected value of 4/2N, 1/2N, 1/2N and 1/2N respectively. The term for three locus association, z'_{ABC} reaches its highest value after only about 0.2N generations; whereas z'_{AC}, that for association between A and C after associations between A and B and between

B and C have been removed, increases together with the marginal associations, z_{AB} and z_{BC} for almost N generations. Subsequently, as only two chromosome types become of higher frequency and eventually are the only ones to remain segregating in the population, knowledge of the frequency of chromosomes carrying the specific alleles A and B and of the frequency of chromosomes carrying alleles B and C gives sufficient information to specify frequencies of chromosomes carrying alleles A and C and alleles A , B and C . Thus the residual likelihood ratio to account for the latter two types of association is zero when any two marginal associations are specified. Of course, when there is no recombination the nominal ordering of loci on the chromosome is irrelevant, so the pairs AB, AC and BC are interchangeable.

Even if there is some recombination among the loci, the quantity z'_{ABC} for three-way association never contributes an appreciable part of the total likelihood ratio after the first few generations. If a pair of the loci are completely linked, e.g., A and B in Figure 2d, z'_{ABC} contributes nothing asymptotically, and very little, or nothing, if the loci are very closely linked (e.g. Figure 2e). The marginal pair-wise associations such as z_{AB} can be obtained from two-locus theory (e.g., Table 1, but for N=8). These always exceed the corresponding conditional associations, e.g., z'_{AB} , by a large amount with tight linkage (Figure 2e) but by very little with looser linkage (Figure 2a). The results of Figure 3 show that with higher L_{AB},L_{BC} values the differences between z_{AB} and z'_{AB} essentially vanish, and z'_{ABC} becomes of similar magnitude to z_{AC} .

Using the results given in Figures 2 and 3, estimates of steady state values of E(z) have been made, and are listed in Table 4.

TABLE 4. Estimates of partitions of E(z) at steady state for three loci (read from Figure 3 for N=80 or Figure 2 for N=20).

L_{AB}	L_{BC}	z_{AB}	z_{BC}	z_{AC}	z'_{AB}	z'_{BC}	z'_{AC}	z'_{ABC}	z_{ABC}
N=80									
8	8	0.032	0.032	0.019	0.032	0.032	0.019	0.012	0.095
4	4	0.059	0.059	0.033	0.057	0.057	0.031	0.014	0.163
2	2	0.10	0.10	0.06	0.10	0.10	0.06	0.02	0.28
N=20									
2	2	0.13	0.13	0.08	0.12	0.12	0.07	0.02	0.35
$\frac{1}{2}$	$\frac{1}{2}$	0.34	0.34	0.22	0.26	0.26	0.14	0.01	0.83
$\frac{1}{4}$	$\frac{3}{4}$	0.55	0.30	0.23	0.41	0.16	0.09	0.00	0.94
0	1	1.00	0.23	0.23	0.77	0.00	0.00	0.00	1.23
$\frac{1}{8}$	$\frac{1}{8}$	0.81	0.81	0.62	0.20	0.20	0.01	0.00	1.63
0	0	1.00	1.00	1.00	0.00	0.00	0.00	0.00	2.00

Since the values of z in successive generations are highly autocorrelated and the number of segregating replicates falls each generation, the estimates shown in Table 4 (and subsequently in Table 5) are based on a visual assessment from the graphs and computer results of where the asymptotic values will lie, taking informally into account the conflict between the number of replicates segregating each generation and the generation number and thus proximity to the asymptote. No standard errors can therefore be attached to the estimates in the Table, but they should act as a guide to the parameter values. The estimates are likely to have lowest precision when L_{AB} and L_{BC} are small since so few replicates are segregating when the steady state values of E(z) are approached. Where, for example, $L_{AB} = L_{BC}$, the same value has been given for $E(z_{AB})$ as for $E(z_{BC})$, regardless of the

actual values in the simulation run, as in Figures 2 and 3; and values of $E(z_{AB})$, for example, obtained in the simulation run are usually given, rather than any two-locus theoretical prediction, so that the partition of $E(z_{ABC})$ is not too disturbed. These values in the table emphasize the small likelihood ratio statistic due to the three-locus association, which always is of small magnitude and only contributes a significant proportion of the total likelihood when values of L_{AB}, L_{BC} are large and the total amount of disequilibrium is small.

Three loci: behaviour of r^2

For comparison, estimates of terms like $E(r^2_{AB})$ and $E(r^2_{ABC})$ at steady state are given in Table 5, based on the same computer runs as Table 4, and using the same estimation procedure.

TABLE 5. Estimates of $E(r^2)$ at steady state for three loci (using same simulation runs as Table 4), and $\alpha = E(D^2_{ABC})/E(p_ap_a,p_bp_b,p_cp_c,)$ at steady state (computed using a moment generating matrix).

L_{AB}	L_{BC}	r^2_{AB}	r^2_{BC}	r^2_{AC}	r^2_{ABC}	Total	α
N=80							
8	8	0.032	0.032	0.018	0.018	0.100	0.018
4	4	0.059	0.059	0.034	0.031	0.183	0.029
2	2	0.09	0.09	0.05	0.05	0.28	0.50
N=20							
2	2	0.12	0.12	0.07	0.07	0.38	0.061
$\frac{1}{2}$	$\frac{1}{2}$	0.33	0.33	0.22	0.37	1.25	0.178
$\frac{1}{4}$	$\frac{3}{4}$	0.51	0.25	0.21	0.37	1.34	0.181
0	1	1.00	0.20	0.20	0.40	1.80	0.187
$\frac{1}{8}$	$\frac{1}{8}$	0.78	0.78	0.60	2.3	4.5	0.436
0	0	1.00	1.00	1.00	4.0	7.0	0.658

At the low values of L_{AB} and L_{BC}, $E(r^2_{ABC})$ is more difficult to estimate than $E(z_{ABC})$ or $E(z'_{ABC})$ because there is much variation in mean level between generations. For $L_{AB} = L_{BC} = 0$, the values roughly agree with those predicted from the transition matrix (Table 3). At high values of L we see that $E(r^2_{AB}) = 1/4L_{AB}$, $E(r^2_{AC}) = 1/4L_{AC} = E(r^2_{ABC})$ approximately, allowing for the fact that recombination fractions used approached 0.1, where map length and recombination fraction do not correspond so closely (Table 1).

While at the low values of L_{AB} and L_{BC} the behaviour of the alternative measures of three-locus association are very different, at higher values this is no longer so. Thus, as L_{AB} and L_{BC} exceed unity, there is little difference between z'_{AC} and z_{AC}, i.e. the likelihood ratio statistic for A and C after fitting AB and BC, or ignoring AB and BC; and the total likelihood ratio statistic z_{ABC} is approximately equal to the equivalent chi-square statistic $(r^2_{AB} + r^2_{AC} + r^2_{BC} + r^2_{ABC})$. The residual d.f. for three locus association, which for the ratio test is z'_{ABC} and for chi-square is r^2_{ABC} (each $\times$ 2N) therefore accounts for roughly the same amount of variation in each case.

The values of r^2 for two locus disequilibrium, e.g. $E(r^2_{AB})$, given in Table 5 are conditional on segregation at all three loci. These differ from those in Table 1 which are for the two-locus model and therefore unconditional on segregation at a third locus. Comparisons between Tables 1 and 5 are also confounded with population size, however. Some three-locus simulation undertaken with N=8 corresponding to the exact transition matrix results given in Table 1 show that there are no quantitatively important differences between the values of $E(r^2_{AB})$ unconditional and conditional on segregation at a third locus at all generations starting from

a population in equilibrium. There may be small differences between the two models, but they could not be detected consistently with the amount of simulation which could be undertaken.

Prediction of $E(r^2_{ABC})$ from ratios of moments

The ratio of moments $E(D^2_{AB})/E(p_a p_{a'} p_b p_{b'})$ is a good predictor of $E(r^2_{AB})$ for two loci, as has been mentioned previously. Equivalent results for the asymptotic value of the ratio $\alpha = E(D^2_{ABC})/E(p_a p_{a'} p_b p_{b'} p_c p_{c'})$ obtained from the first eigenvector of the appropriate moment generating matrix (Hill, 1974b) are given in Table 5. The correspondence between α and $E(r^2_{ABC})$ is seen to be very poor at low values of L_{AB} and L_{BC}, and since α is essentially independent of population size for given L_{AB}, L_{BC} the departure would become greater as N is increased. However for L_{AB}, $L_{BC} > 1$ or so, there is seen to be reasonable agreement. This is not very useful to us however, for the simple approximation $E(r^2_{ABC}) = 1/4L_{AC}$ is also satisfactory.

Four or more loci

In view of the computer time which would be necessary in each replicate run to enable a complete partition of the total likelihood ratio, only the total value has been computed for some parameter sets with four and five loci. Results, as $E(z_{ABCD})$ and $E(z_{ABCDE})$ which are $1/2N \times$ the total likelihood ratio statistics, are given for four and five loci in Figures 4 and 5, respectively. The computational problems of fixation of most replicates long before the steady state is reached is even more acute with these higher numbers of loci, so results are only approximate at high generation number. Indeed, when the total map length is zero, the steady state

has clearly not been reached in the simulation, as shown by comparison of Figure 4 with Table 3. The total likelihood ratio (× 1/2N) can be compared using Figures 2, 4 and 5 for the outermost loci a specific distance apart, but with different numbers of intervening loci included. With L=4 between the outside pair, the asymptotic values of z are 0.08, 0.35, 0.8 and 1.6 approximately, for 2, 3, 4 and 5 loci with N=20 . In Figure 4 alternative configurations of distances among the loci, for given distance between the outside pair, are given for $L_{AD} = 1$.

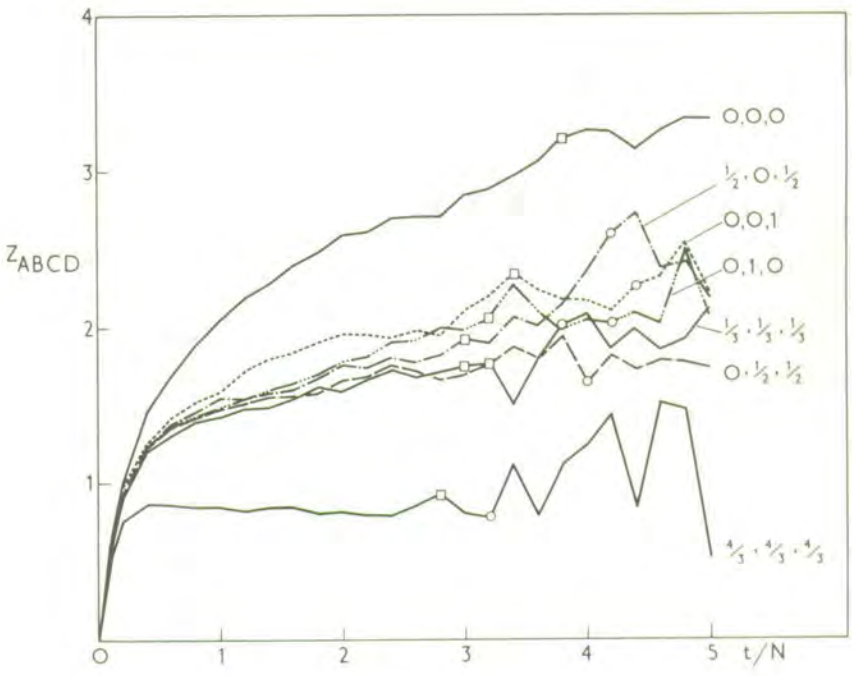

FIGURE 4. *As Figure 1, but* z_{ABCD} *(total likelihood ratio/2N) for four loci, in each case with N=20 .*

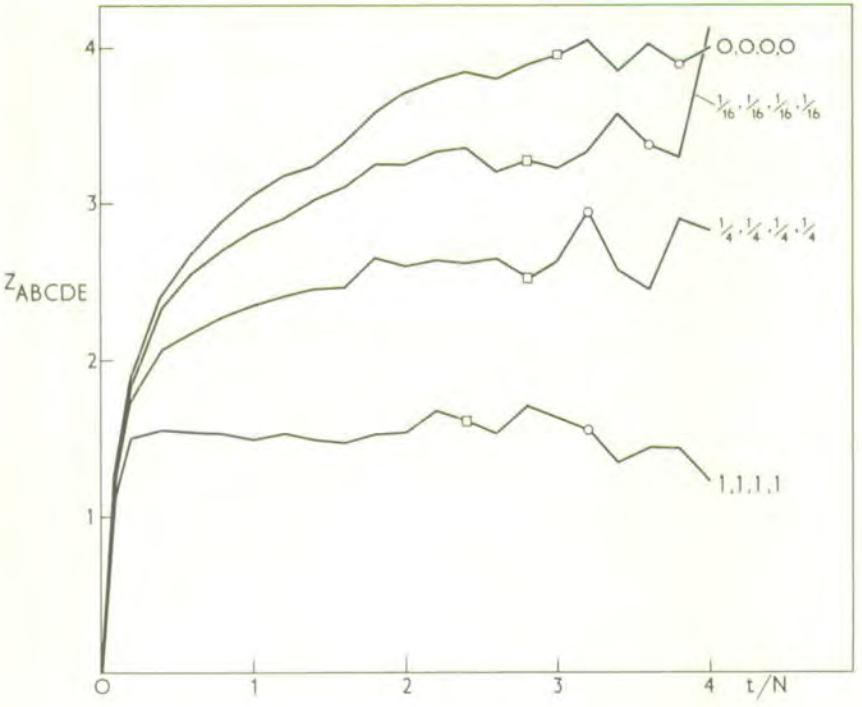

FIGURE 5. *As Figure 1, but z_{ABCDE} (total likelihood ratio/2N) for five loci, in each case with N=20 .*

The effect on the total value of z is not large relative to changes in the value of L_{AD} , as also found for three loci (Figure 2, Table 4). A bigger contrast would be found, however, for larger values of L_{AD} between say $L_{AB} = L_{BC} = L_{CD} = L_{AD}/3$ and $L_{AB} = L_{BC} = 0$, $L_{CD} = L_{AD}$. The general pattern of the four and five locus results is similar to that for three loci.

For values of L in excess of about unity between all adjacent pairs of loci, predictions of the chi-square statistics

(or quantities like r^2) can be obtained by extending the arguments used earlier (eq.19). These suggest that for four equally spaced loci the expected value of the total chi-square statistic with 11 d.f. is $2N \times 19/12L$, where L is the distance between adjacent loci.

DISCUSSION

Statistical

Our results have been presented solely in terms of chi-square or likelihood ratio statistics, usually with a scalar multiplier $(1/2N)$. These non-negative quantities may be adequate for a discussion of drift at neutral loci, where the expected values of disequilibria are zero, but are less so for discussing selection or migration where the signs of the disequilibria may be important. As Lewontin (1974) and others have pointed out, with selection as the main cause of disequilibrium, one might expect it to be of the same sign and magnitude in different populations. With data on animal or plant populations, however, the experimentalist or field worker is likely to test for association (disequilibrium) using standard chi-square or likelihood ratio methods. If he finds evidence for the presence of such disequilibria its sign and magnitude can then be calculated, and expressed in terms of, for example, D_{AB}, r_{ABC} or the quantity which has to be added to the chromosome frequencies in order to satisfy Bartlett's criterion (14). With neutrality, all such quantities will be distributed about zero. Of course, with symmetric selection models in infinite population disequilibria of the same magnitude but opposite sign can occur, the value found in any particular population depending on the initial frequencies (Bodmer and Felsenstein, 1967).

More use has been made here of likelihood ratio than chi-square statistics. When there is little departure from random association, the two methods give essentially the same values, and both are distributed as χ^2 distribution in large samples from equilibrium populations. With very large departures from random association they behave rather differently, especially when one type is very rare. For example, if in a sample of 100 chromosomes the gene frequency at each of four loci is 0.1, the rarest type has an expected frequency of 0.0001 or expected number of occurrences of 0.01 . If such a chromosome is actually obtained it contributes $(1-0.01)^2/0.01 \sim 100$ to chi-square, but only $2\log(1/0.01) = 9.2$ to the likelihood ratio. With field data, such rare classes might be pooled, but when making predictions of behaviour and making analyses over many populations as done here, arbitrary decisions to pool classes would have an uncertain effect on the results. Furthermore, pooling of classes with field data would cause some loss of information, and difficulties would clearly arise with the partition of the chi-square or likelihood ratio. In practice the most satisfactory procedure appears to be to carry out a complete partititoned likelihood ratio analysis before contemplating pooling of classes. With small numbers in the sub-classes appropriate exact tests seem to be required and some work on these is in progress. An extension of the analysis for data on diploid individuals including partition of the likelihood for several loci has been given by Hill (1975) using maximum likelihood methods and assuming random mating populations.

There remain unresolved problems in this area of analyzing multi-dimensional contingency table data, and we have already considered alternative forms of the partition. Genetic

interpretations of, say, three locus associations measured by Bartlett's criterion are hard to visualize, and a fuller attempt is given elsewhere (Hill, 1975). We would not have to concern ourselves with such statistical problems of specifying three locus associations if the alternative methods did not give such radically different answers. But, as we have seen, with very tight recombination, the parameters z'_{ABC} (essentially Bartlett's criterion) and r^2_{ABC} (essentially Lancaster's criterion) approach 0 and 2log2N-4 respectively. For values of L(N x population size) much in excess of unity the criteria do not differ so greatly.

Genetical

The objective of this work has been to study the process of drift for several neutral genes, so that population behaviour under the neutrality hypothesis can be determined. This is only a start, in that one has also to look at behaviour with selection and the other evolutionary forces of migration, mutation and population structure or non-random mating. Whilst there is already much information on two loci with selection, that on three or more is much more limited. In the best known study, that of Franklin and Lewontin (1970), simulation was used so there was some confounding of population size and selection effects. They concluded that, with a model of linked genes with heterozygote superiority at each locus and multiplicative effects over loci, the population would tend to only two or a few complementary chromosome types. This could correspond to the state of a population with neutral, but closely linked genes, after all but these few chromosome types had been lost by chance. With selection and heterozygote superiority, segregation at all, or most of the loci would be expected in all populations, and the indi-

vidual gene frequencies would tend to be intermediate. Neither of these conditions would be expected with neutrality. With selection in infinite populations, however, very many equilibrium chromosome frequency sets are possible (Feldman et al., 1974) a proportion of which are stable (Karlin, 1975). Thus, as a consequence of founder effects it would be possible for different chromosome polymorphisms to be segregating among populations. It is difficult to differentiate between migration and selection as forces maintaining similarity of frequencies in different populations. There remain problems in distinguishing between the alternative models.

With the exception of the limiting case of $L \to 0$ and for $L > 1$ we have no explicit solutions. However our observation is that, among populations segregating at three loci, the contribution (z'_{ABC}) to the likelihood ratio made by three locus association is always small in magnitude, although becoming a significant part of the total when the product of map distance between the extreme loci on the chromosome and population size is large. This may well be the most useful result to come from this study. Having established what happens with neutral genes we now require detailed information on the equivalent predictions of these likelihood ratio parameters for models of selection and other forces.

Even with two loci, little is known about the joint behaviour of drift and selection; workers have tended to study either drift with neutral genes or selection in infinite populations. It is unlikely that many analytical results will be obtained to such involved equations, but simulation of at least some parameter sets should be feasible. An illustration of the relevance of such a study is that a neutral gene as defined by Kimura and Ohta (1971) is one with a selective value of less than about $1/N$, i.e. neutrality is defined in

terms of population size. What then are expected disequilibria for pairs or more of loci with selective values of order 1/N ? The little evidence available for two loci suggests that r^2_{AB} is similar with such selective values for heterotic loci having fitnesses combined multiplicatively to that for neutral genes (Hill and Robertson, 1968), but the results with epistasis or with more loci may be very different. The tendency may be towards a reduction in the incidence of very rare chromosome types involving several loci.

For two locus disequilibria, the critical values of recombination fraction or map length are of order 1/N : i.e., $r^2_{AB} \sim 1/(4Ny_{AB}+1)$, using Sved and Feldman's (1973) formula, and just $1/4Ny_{AB}$ for $Ny_{AB} > 1$ approximately. Thus, if Ny_{AB} exceeds unity by an order of magnitude, r^2_{AB} is very small, and for appreciable disequilibrium to be found, populations must be of small effective size and/or linkage must be tight. Regrettably, but not surprisingly, perhaps, we find the same dependency for three loci, except that the three locus association (from the likelihood ratio) is always small. For Ny (or L) values in excess of unity, the total three locus association is also roughly proportional to 1/Ny . Our simulations have been based on constant population size, however, and a small founder population could induce considerable disequilibrium for long periods thereafter, even in a large population.

As yet our discussion of the relative effects of neutrality versus selection on multi-locus disequilibrium remains somewhat inconclusive. Is there then any point in an experimentalist or field worker estimating disequilibrium at all? Undoubtedly there is, for if he has already collected information on genotype frequencies, estimates of disequilibrium can always be made from the same data using relevant statistical

techniques, and it will provide results on which this and any subsequent theory can be tested.

ACKNOWLEDGEMENTS

I am indebted to Mrs. Marjorie McEwan for much assistance, with both computer programming and presentation of results.

REFERENCES

Bartlett, M.S. (1935). J.R. Statist. Soc. Suppl. 2: 248-252.

Bennett, J.H. (1954). Ann. Eugen. 18: 311-317.

Bodmer, W.F. and J. Felsenstein. (1967). Genetics 57: 237-265.

Feldman, M.W., I. Franklin and G.J. Thomson. (1974). Genetics 76: 135-162.

Fienberg, S.E. (1970). Ecology 51: 419-433.

Fisher, R.A. (1930). The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Franklin, I. and R.C. Lewonton. (1970). Genetics 65: 707-734.

Geiringer, H. (1944). Ann. Math. Stat. 15: 25-57.

Goodman, L.A. (1969). J.R. Statist. Soc. B 31: 486-498.

Goodman, L.A. (1970). J. Amer. Statist. Assoc. 65: 226-256.

Hill, W.G. (1969). Jap. J. Genet. 44 (Suppl.1): 144-151.

Hill, W.G. (1974a). Theor. Pop. Biol. 5: 366-392.

Hill, W.G. (1974b). Theor. Pop. Biol. 6: 184-198.

Hill, W.G. (1974c). Heredity 33: 229-239.

Hill, W.G. (1975). Biometrics (in press).

Hill, W.G. and A. Robertson. (1966). Genet. Res. 8: 269-294.

Hill, W.G. and A. Robertson. (1968). Theor. Appl. Genet. 38: 226-231.

Karlin, S. (1975). Theor. Pop. Biol. (in press).

Kimura, M. (1955). Evolution 9: 419-435.

Kimura, M. and T. Ohta. (1971). Theoretical Aspects of Population Genetics. Princeton Univ. Press, Princeton, N.J.

Lancaster, H.O. (1951). J.R. Statist. Soc. B 13: 242-249.

Lewontin, R.C. (1964a). Genetics 49: 49-67.

Lewontin, R.C. (1964b). Genetics 50: 757-782.

Lewontin, R.C. (1974). The Genetic Basis of Evolutionary Change. Columbia Univ. Press, New York.

Ohta, T. and M. Kimura. (1969). Genet. Res. 13: 47-55.

Plackett, R.L. (1962). J.R. Statist. Soc. B 24: 162-166.

Slatkin, M. (1972). Genetics 72: 157-168.

Smouse, P.E. (1974). Genetics 76: 557-565.

Sokal, R.F. and F.J. Rohlf. (1969). Biometry. Freeman, San Francisco.

Strobeck, C. (1973). Genet. Res. 22: 195-200.

Sved, J.A. (1971). Theor. Pop. Biol. 2: 125-141.

Sved, J.A. and M.W. Feldman. (1973). Theor. Pop. Biol. 4: 129-132.

Wright, S. (1933). Proc. Nat. Acad. Sci. U.S. 19: 420-433.

28

Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population.

by

William G. Hill

Reprinted from THEORETICAL POPULATION BIOLOGY
Vol. 8, No. 2, October 1975
Printed in Belgium

Linkage Disequilibrium among Multiple Neutral Alleles Produced by Mutation in Finite Population

WILLIAM G. HILL

Institute of Animal Genetics, West Mains Road, Edinburgh, EH9 3JN, Scotland

Received September 20, 1974

An analysis is undertaken for a finite random mating population of the linkage disequilibrium between two loci, at both of which all alleles are neutral, all mutant alleles differ from existing ones and several may be segregating at any time. Formulae are derived for the expected total squared disequilibrium, measured as the sum of squares of disequilibria between all pairs of alleles. The ratio of this quantity to the expected value of the product of the heterozygosities at the two loci is similar to that obtained previously by Ohta and Kimura for two nucleotide sites at each of which not more than two mutant types can segregate at any time.

INTRODUCTION

A model of mutation was introduced by Kimura and Crow (1964) in which each mutant allele at a locus is assumed never to have existed previously in the population, and several may be segregating at any time. This has become known as the "infinite alleles" model and was used by Kimura and Crow (1964) to compute expected homozygosities and the effective number of segregating alleles in a finite population, assuming all alleles had no effect on fitness. An alternative model is one of many nucleotide sites at which there are molecular mutations, with mutation at each site so rare that new mutants only occur at nonsegragating sites and not more than two mutants are present at any site at one time. This "infinite sites" model, originally of Karlin and McGregor (1967), was used by Kimura (1969) to find the expected number of heterozygous sites in finite population. Ewens (1974) has recently contrasted the two models.

Using the infinite sites model, Ohta and Kimura (1969) computed the variance of linkage disequilibrium between pairs of sites. In this note the equivalent result is obtained for the infinite alleles model in which account has now to be taken of several alleles segregating at each locus. The results, however, can be expressed in a simple way.

Analysis

The following definitions and assumptions are made:

t is the generation number;
p_h , p_i are the frequencies of alleles A_h , A_i at the A locus;
q_j , q_k are the frequencies of alleles B_j , B_k at the B locus;
f_{ij} is the frequency of the chromosome A_iB_j ;
$D_{ij} = f_{ij} - p_iq_j$ is the disequilibrium between alleles A_i and B_j ;
α and β are the number of alleles at loci A and B, respectively, which have existed by generation t;
N is the number of monecious diploids in the population, which is assumed to be random mating;
u and v are the mutation rates per generation at loci A and B, respectively, with all mutations being to new alleles;
c is the recombination fraction between loci A and B; for simplicity N is assumed to be sufficiently large and u, v and c sufficiently small that $1/N^2$, u^2, v^2 and c^2 can be ignored relative to $1/N$, i.e., u, v and c are $O(1/N)$, but the effects of relaxing this assumption are discussed;
$U = Nu$, $V = Nv$, $C = Nc$.

A haploid model is used in which mutation and recombination are assumed to change chromosome frequencies deterministically, and from these new frequencies a sample of $2N$ chromosomes is taken from a multinomial distribution. All existing and mutant alleles are assumed to be neutral, so no frequency changes occur from selection, and the expected values of all disequilibria are zero.

Let us consider the vector $\mathbf{y}_{hi,jk(t)}$ of moments associated with alleles A_h , A_i , B_j and B_k at generation t, defined by

$$\mathbf{y}_{hi,jk(t)} = \begin{pmatrix} E(p_hp_iq_jq_k)_{(t)} \\ E(p_hq_jD_{ik} + p_hq_kD_{ij} + p_iq_jD_{hk} + p_iq_kD_{hj})_{(t)} \\ E(D_{hj}D_{ik} + D_{hk}D_{ij})_{(t)} \end{pmatrix},$$

where $E(\ )_{(t)}$ denotes the expected value of the quantity at generation t. Denoting by $\mathbf{D}$, $\mathbf{R}$ and $\mathbf{M}$ the transition matrices for changes in these moments due to drift, recombination and mutation, respectively, we have for the haploid model

$$\mathbf{y}_{hi,jk(t+1)} = \mathbf{DRM}\mathbf{y}_{hi,jk(t)} . \tag{1}$$

Using, for example, the methods of Hill (1974), it can be shown that

$$\mathbf{D} = \begin{pmatrix} (1-z)^2 & z(1-z)^2 & z^2(1-z) \\ 0 & (1-z)(1-2z)^2 & 2z(1-z)(1-2z) \\ 2z(1-z) & 2z(1-z)^2 & (1-z)[z^2 + (1-z)^2] \end{pmatrix}, \tag{2}$$

where $z = (2N)^{-1}$. The matrix $\mathbf{D}$ is essentially that obtained for two alleles at each locus by Hill and Robertson (1968) and subsequently others, and has been given recently by Weir and Cockerham (1974), for a vector $\mathbf{y}$ defined slightly differently, in the more general multiple allele case. From (2),

$$\mathbf{D} = \mathbf{I} - (1/N)\begin{pmatrix} 1 & -1/2 & 0 \\ 0 & 5/2 & -1 \\ -1 & -1 & 3/2 \end{pmatrix} + O(N^{-2}),$$

where $\mathbf{I}$ is the identity matrix. The recombination and mutation matrices are, respectively,

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-c & 0 \\ 0 & 0 & (1-c)^2 \end{pmatrix} = \mathbf{I} - (1/N)\begin{pmatrix} 0 & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & 2C \end{pmatrix} + O(N^{-2}),$$

and

$$\mathbf{M} = (1-u)^2(1-v)^2\mathbf{I} = \mathbf{I} - (1/N)(2U + 2V)\mathbf{I} + O(N^{-2}).$$

Therefore,

$$\begin{aligned} \mathbf{DRM} &= \mathbf{I} - (1/N)\begin{pmatrix} 1+2U+2V & -1/2 & 0 \\ 0 & 5/2+C+2U+2V & -1 \\ -1 & -1 & 3/2+2C+2U+2V \end{pmatrix} \\ &\quad + O(N^{-2}) \\ &= \mathbf{I} - (1/N)\mathbf{P} + O(N^{-2}), \end{aligned} \tag{3}$$

say.

Now consider variation from all alleles present at the two loci and define

$$\mathbf{x}_{(t)} = \sum_{h\neq i}^{\alpha}\sum \sum_{j\neq k}^{\beta}\sum \mathbf{y}_{hi,jk(t)}.$$

Let us denote by $H_A = \sum\sum_{h\neq i}^{\alpha} p_h p_i$ and $H_B = \sum\sum_{j\neq k}^{\beta} q_j q_k$ the heterozygosities at the A and B loci. Also, since

$$\sum_{i=1}^{\alpha} D_{ij} = \sum_{j=1}^{\beta} D_{ij} = 0,$$

we have

$$\sum_{h\neq i}^{\alpha}\sum_{k\neq j}^{\beta} D_{hk} = D_{ij}.$$

Using these relationships, $\mathbf{x}_{(t)}$ can be rewritten into more meaningful quantities:

$$\mathbf{x}_{(t)} = \begin{pmatrix} E(H_A H_B)_{(t)} \\ 4E\left(\sum_i \sum_j p_i q_j D_{ij}\right)_{(t)} \\ 2E\left(\sum_i \sum_j D_{ij}^2\right)_{(t)} \end{pmatrix}. \tag{4}$$

Using (1) and (4),

$$\mathbf{x}_{(t+1)} = \mathbf{DRM}\mathbf{x}_{(t)} + \mathbf{w}_{(t)}, \tag{5}$$

and substituting in (3),

$$\mathbf{x}_{(t+1)} = [\mathbf{I} - (1/N)\mathbf{P}]\,\mathbf{x}_{(t)} + \mathbf{w}_{(t)} + O(N^{-2}), \tag{6}$$

where $\mathbf{w}_{(t)}$ denotes the vector of expected increments due to mutation to new alleles.

Before considering the magnitude of $\mathbf{w}_{(t)}$ it is helpful to review the analysis of Kimura and Crow (1964) for single loci. Then $E(H_A)$ is reduced by a factor $(1 - 1/2N)(1 - u)^2 = 1 - 1/2N - 2u + O(N^{-2})$ due to drift and mutation from old alleles, and increased by the quantity (expected number of mutants $\times$ heterozygosity from a new mutant) which equals $2Nu \times 2/2N = 2u$, again ignoring terms in N^{-2}. Hence,

$$E(H_{A(t+1)}) = E(H_{A(t)})(1 - 1/2N - 2u) + 2u,$$

and at equilibrium

$$E(H_A) = 4Nu/(4Nu + 1) = 4U/(4U + 1), \tag{7}$$

corresponding to a value of $1/(4U + 1)$ for the homozygosity, derived by Kimura and Crow.

The increment in $E(H_A H_B)$ due to new mutation is, by extending the single locus arguments,

$$w_{1(t)} = 2vE(H_{A(t)}) + 2uE(H_{B(t)}) + O(N^{-2}).$$

The values of disequilibria associated with a new mutant, $A_{\alpha+1}$, at the A locus, occurring on a chromosome containing B_j in some population, are

$$D_{\alpha+1,j} = 1/2N - q_j/2N = (1 - q_j)/2N,$$
$$D_{\alpha+1,k} = -q_k/2N, \qquad k \neq j.$$

The probability that the mutation is associated with B_j is q_j ; hence the total increment in $\sum\sum D_{ij}^2$ in this population due to the single mutation at the A locus is expected to be

$$(2N)^{-2}\sum_j q_j\left[(1-q_j)^2+\sum_{k\neq j} q_k{}^2\right]=(2N)^{-2}\left(1-\sum q_j{}^2\right)=(2N)^{-2}H_B\,.$$

The expected increment in $E(\sum\sum_{ij} D_{ij}^2)$ over all populations is thus $2Nu(2N)^{-2}H_B$, and is $O(N^{-2})$, as is that due to mutation at B. A similar argument holds for the element $w_{2(t)}$ which is also $O(N^{-2})$. Therefore,

$$\mathbf{w}'_{(t)}=(2vH_{A(t)}+2uH_{B(t)}\,,0,0)+O(N^{-2}).$$

Using (7), the steady flux value of $\mathbf{w}_{(t)}$ is

$$\mathbf{w}'=(8UV/N)((4U+1)^{-1}+(4V+1)^{-1},0,0), \tag{8}$$

and using (6), the steady flux value of $\mathbf{x}_{(t)}$ is

$$\mathbf{x}=N\mathbf{P}^{-1}\mathbf{w}. \tag{9}$$

From (3), (8) and (9),

$$\begin{aligned}\mathbf{x}=8UV[(4U+1)^{-1}+(4V+1)^{-1}][9+26C+54(U+V)+8C^2\\+76C(U+V)+80(U+V)^2+16C^2(U+V)+48C(U+V)^2+32(U+V)^3]^{-1}\\\times\begin{pmatrix}11+26C+32(U+V)+8C^2+24C(U+V)+16(U+V)^2\\4\\10+4C+8(U+V)\end{pmatrix},\end{aligned} \tag{10}$$

where $E(H_AH_B)=x_1$, $E(\sum\sum p_iq_jD_{ij})=\frac{1}{4}x_2$ and $E(\sum\sum D_{ij}^2)=\frac{1}{2}x_3$.

Steady flux values of $E(H_AH_B)$ and $E(\sum\sum D_{ij}^2)$, together with functions of them, are given in Table I for some examples of U, V and C. These results, obtained from (6), (9) and (10) rest on the assumption that u, v and c are $O(N^{-1})$ and N is large. In particular, removal of the restriction of c to small values would be desirable. Equation (5) can be used directly, to give

$$\bar{\mathbf{x}}=(\mathbf{I}-\mathbf{DRM})^{-1}\mathbf{w}, \tag{11}$$

but an explicit form for $\bar{\mathbf{x}}$ has not been obtained, although presumably could be after much manipulation (cf. Littler, 1973). The approximation seems satisfactory for most purposes however: for example, with $N=100$ and $U=V=0.01$ the values of $E(\sum\sum D_{ij}^2)$ obtained using (11) are 0.06281, 0.02390, 0.00366, 0.00165 and 0.00098 compared with the values obtained from (10) of 0.06273, 0.02369, 0.00345, 0.00144 and 0.00073 for $C=0.1$, 1, 10, 25 and 50 respectively, i.e., recombination fractions up to 0.5, where the disequilibrium is trivial.

TABLE I

Computation of Heterozygosities and Disequilibria for a Range of Parameters (see text for definitions)

U	V	$E(H_A)$	$E(H_B)$	$E(H_A)E(H_B)$	C	$E(H_AH_B)$	$\frac{E(H_AH_B)}{E(H_A)E(H_B)}$	$E(\Sigma\Sigma D_{ij}^2)$	σ_d^{2*}
$\to 0$	$\to 0$	$4U$	$4V$	$16UV$	0	$19.56UV$	1.222	$8.89UV$	0.455
					0.1	$18.73UV$	1.171	$7.12UV$	0.380
					1	$16.74UV$	1.047	$2.60UV$	0.156
					10	$16.03UV$	1.002	$0.37UV$	0.023
$\to\infty$	$\to\infty$	1	1	1	C	1	1	0	0
U	V	$\frac{4U}{4U+1}$	$\frac{4V}{4V+1}$	†	$\to\infty$	†	1	0	0
				$\times 100$		$\times 100$		$\times 100$	
0.01	0.01	0.038	0.038	0.1479	0	0.1772	1.198	0.0773	0.436
					0.1	0.1708	1.154	0.0627	0.367
					1	0.1544	1.043	0.0237	0.153
					10	0.1482	1.002	0.0035	0.023
0.1	0.1	0.286	0.286	8.1633	0	8.865	1.086	2.850	0.322
					0.1	8.753	1.072	2.477	0.283
					1	8.374	1.026	1.149	0.137
					10	8.174	1.001	0.187	0.023
0.01	0.19	0.038	0.432	1.6608	0	1.804	1.086	0.580	0.322
					0.1	1.781	1.072	0.504	0.283
					1	1.704	1.026	0.234	0.137
					10	1.663	1.001	0.038	0.023
1	1	0.8	0.8	64.000	0	64.185	1.003	6.003	0.094
					0.1	64.175	1.003	5.783	0.090
					1	64.116	1.002	4.352	0.068
					10	64.015	1.000	1.258	0.020

† $E(H_AH_B) = E(H_A)E(H_B)$.

Discussion

Let us contrast our result (10) with that obtained by Ohta and Kimura (1971) for the multiple site model. Not more than two types can be segregating at either of the two sites at any time, so let the frequency of type A_i at the first site be p

and of type B_j at the second be q and D the disequilibrium between A_i and B_j. Ohta and Kimura found (their Eqs. 7 and substituting our C for their R)

$$\begin{aligned} E[p(1-p)\,q(1-q)] &= N_e K(11 + 26C + 8C^2 + 2/N)/(9 + 26C + 8C^2) \\ E[(1-2p)(1-2q)D] &= 4N_e K(1 + 1/N)/(9 + 26C + 8C^2) \\ E[D^2] &= N_e K(1 + 1/N)(5 + 2C)/(9 + 26C + 8C^2), \end{aligned} \tag{12}$$

where N_e is the effective population size, used also to define C, and $K = v_s/[4N(\log_e 2N + 1)]$ where v_s is "the number of pairs of nucleotide sites that start segregating simultaneously in the entire population each generation, considering only those pairs of sites that are separated by a distance corresponding to a recombination fraction c." In Ohta and Kimura's model, N is the population size when the mutant occurs at the site which is not previously segregating. With only two alleles segregating at a locus in our model,

$$D_{11} = -D_{12} = -D_{21} = D_{22} = D,$$

say,

$$H_A H_B = 4p(1-p)\,q(1-q), \qquad 4\sum\sum p_i q_j D_{ij} = 4(1-2p)(1-2q)$$

and

$$2\sum\sum D_{ij}^2 = 8D^2.$$

If U and V are both small relative to unity, the expected heterozygosity is small at each locus and we can assume only two alleles are segregating; (10) gives

$$\begin{aligned} E[p(1-p)\,q(1-q)] &= \tfrac{1}{4}x_1 = 4UV(11 + 26C + 8C^2)/(9 + 26C + 8C^2) \\ E[(1-2p)(1-2q)D] &= \tfrac{1}{4}x_2 = 16UV/(9 + 26C + 8C^2) \\ E[D^2] &= \tfrac{1}{8}x_3 = 4UV(5 + 2C)/(9 + 26C + 8C^2). \end{aligned} \tag{13}$$

These Eqs. (13) are the same as (12) above of Ohta and Kimura, providing a term of $1/N$ is ignored relative to unity (as it is in the rest of their diffusion analysis) and with $4UV$ replacing $N_e K = (N_e/N)\,v_s/[4(\log_e 2N + 1)]$. Both $4UV$ and $N_e K$ are proportional to the number of new mutant alleles (types) occurring simultaneously in any generation at the two loci (sites), but differ by a scalar multiplier. Of these quantities UV seems more tangible than v_s; although U or V are products of mutation rate and population size, and thus never estimated directly without full past knowledge of the population, they can be estimated from the marginal heterozygosities H_A and H_B.

The quantity σ_d^2, the squared "standard linkage disequilibrium" given by

$$\sigma_d^2 = E[D^2]/E[p(1-p)\,q(1-q)]$$

was also discussed by Ohta and Kimura (1971). Both in their infinite site model (again ignoring terms of order N^{-1}) and in this infinite allele model with U and V small,

$$\sigma_d^2 = (5 + 2C)/(11 + 26C + 8C^2) \tag{14}$$

from (12) and (13), and σ_d^2 approaches 5/11 if $Nc(=C)$ is small and $1/4Nc$ if Nc is large. A multiple allele equivalent to σ_d^2, say σ_d^{2*}, can be defined by

$$\sigma_d^{2*} = E\left(\sum\sum D_{ij}^2\right)\Big/E(H_A H_B)$$
$$= \tfrac{1}{2}x_3/x_1 .$$

This reduces to σ_d^2 when not more than two alleles are segregating at each locus, and is then given by (14). Thus for very small values of U and V (i.e., population size x mutation rate) the values of σ_d^{2*} are the same as in the infinite sites model. With more mutation $E(H_A H_B)$ increases faster than $E(\sum\sum D_{ij}^2)$ and values of σ_d^{2*} are smaller (Table I); the change in σ_d^{2*} with an increase of U and V from 0.01 to 0.1, corresponding to heterozygosities increasing from 0.038 to 0.286, is, however, only about 25% (from 0.436 to 0.322) for $C = 0$, about 10% for $C = 1$ and is negligible for $C = 10$. As shown by (10) and Table I, σ_d^{2*} is a function of $U + V$ and not of U or V separately.

As a consequence of linkage the expectation of the product of the heterozygosities, $E(H_A H_B)$, exceeds the product of their expectation, i.e., they are correlated. A measure of their association, $E(H_A H_B)/[E(H_A)\,E(H_B)]$ is shown in Table I, but does not exceed unity by more than 22%. For large C, $E(H_A H_B)$ given by x_1 in (10) approaches $16UV(4U + 1)^{-1}(4V + 1)^{-1} = E(H_A)\,E(H_B)$, and is always close to this value for $C \geqslant 10$.

The model of Crow and Kimura (1964) of an infinite number of neutral alleles has been criticised in that it predicts too many segregating alleles in large populations and that the observed range of heterozygosity in nature corresponds to a very narrow range, say 0.015 to 0.057, of population size $\times$ recombination values (see e.g. Lewontin, 1974). Ohta and Kimura (1973) proposed a new model of electrophoretically detectable alleles, which gave a predicted heterozygosity of $1 - (1 + 8U)^{-1/2}$ rather than $1 - (1 + 4U)^{-1}$ as in (7). Although heterozygosities approach unity at much higher values of U with this model, the range over which heterozygosity is very sensitive to changes in U is not greatly affected. If the new model were incorporated into the analysis of this paper the total disequilibrium would be reduced at higher values of U and V, but will only change in proportion to the heterozygosities so σ_d^{2*} is unlikely to be substantially affected.

There are several alternative ways of describing multiple-allele linkage disequilibria, that used here ($\sum\sum D_{ij}^2$) is one of the simplest but perhaps too

great a condensation of the information, certainly if the D_{ij} do not have a mean of zero. In an analysis of chromosomes taken from a population the standard test for disequilibrium would be by chi-square in a two-way contingency table. The expected frequencies in samples of size n are np_iq_j and the observed frequencies are nf_{ij}, so the chi-square statistic is $n\sum\sum (D_{ij}^2/p_iq_j)$, with degrees of freedom dependent on the number of alleles segregating. With two alleles at each locus this statistic equals $nD^2/p(1-p)\,q(1-q)$ with 1 df (Hill and Robertson, 1968). The moment formulation gives the ratio of expectations of numberator and denominator, which approximates the required expectation of the ratio where it has been examined (Ohta and Kimura, 1969; Hill, unpublished), and Littler (1973) discusses the conditions where this is likely to occur. The same simplification appears to hold less well with multiple alleles, but analysis of the behaviour of the chi-square statistic requires Monte Carlo simulation.

There has been little theoretical study yet of disequilibrium between multiple alleles at loci which have an effect on fitness, yet such disequilibrium can occur as in the *HL-A* system of man (e.g., Cavalli-Sforza and Bodmer, 1971, Section 5.11). Thus comparisons between the predictions from neutral and selective models can not be made at this stage.

References

Cavalli-Sforza, L. L. and Bodmer, W. F. 1971. "The Genetics of Human Populations," Freeman, San Francisco.

Ewens, W. J. 1974. A note on the sampling theory for infinite alleles and infinite sites models, *Theor. Pop. Biol.* **6**, 143–148.

Hill, W. G. 1974. Disequilibrium among several linked neutral genes in finite population. II. Variances and covariances of disequilibria, *Theor. Pop. Biol.* **6**, 184–198.

Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations, *Theor. Appl. Genet.* **38**, 226–231.

Karlin, S. and McGregor, J. L. 1967. The number of mutant forms maintained in a population, *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **4**, 415–438.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics* **61**, 893–903.

Kimura, M. and Crow, J. F. 1964. The number of alleles that can be maintained in a finite population, *Genetics* **49**, 725–738.

Lewontin, R. C. 1974. "The Genetic Basis of Evolutionary Change," Columbia University Press, New York.

Littler, R. A. 1973. Linkage disequilibrium in two-locus, finite, random mating models without selection or mutation, *Theor. Pop. Biol.* **4**, 259–275.

Ohta, T. and Kimura, M. 1969. Linkage disequilibrium due to random genetic drift, *Genet. Res.* **13**, 47–55.

Ohta, T. and Kimura, M. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population, *Genetics* **68**, 571–580.

29

Estimation of linkage disequilibrium in randomly mating populations

by

William G. Hill

Reprinted from HEREDITY, Volume 33, Part 2, pp. 229-239, October 1974

ESTIMATION OF LINKAGE DISEQUILIBRIUM IN RANDOMLY MATING POPULATIONS

WILLIAM G. HILL
Institute of Animal Genetics, Edinburgh EH9 3JN

Received 15.xi.73

SUMMARY

The degree of linkage disequilibrium, D, between two loci can be estimated by maximum likelihood from the frequency of diploid genotypes in a sample from a random-mating population. Haploid genotypes can be identified directly in some species from a sample of chromosomes extracted from the population and made homozygous, or by test crossing. The maximum likelihood estimators of D are described, with examples, for both methods, including the cases where both loci are codominant and one or both are dominant.

The efficiencies of the methods are compared when $D = 0$: If both loci are codominant the estimate of D has the same variance,

$$V(\hat{D}) = p(1-p)q(1-q)/N,$$

from a sample of N identified diploids as from N identified haploid types, where p and q are the gene frequencies; therefore the diploid method is more efficient in practice since less labour is required. With dominance at either locus $V(\hat{D})$ is lower for samples of the same size using the haploid method if the dominant alleles are at high frequency.

1. INTRODUCTION

Now that it is possible to use starch gel electrophoresis to type the same individual for several different polymorphic loci, some of which may be linked, associations between the frequencies of alleles at two or more loci are being studied. Allard and his group with plants (*e.g.* Allard, Babbel, Clegg and Kahler, 1972), several groups with *Drosophila* (Prakash and Lewontin, 1968, 1971; Kojima, Gillespie and Tobari, 1970; Zouros and Krimbas, 1972; Charlesworth and Charlesworth, 1973; Franklin, 1973) and Webster (1973) in salamander have found such linkage disequilibrium, although in the *Drosophila* cases usually associated with a chromosomal inversion. Mukai, Mettler and Chigusa (1971) however, did not find any associations among linked genes in *D. melanogaster*. Sinnock and Sing (1972*a*, *b*) found some evidence of disequilibrium among loci in man, but these loci were not known to be linked. A group in this laboratory (D. A. Briscoe, J. M. Malpica and A. Robertson) are also doing similar analyses on *Drosophila* populations which will be reported subsequently. In view of the number of these studies being undertaken, whatever their possible contribution to population genetics, it seems worth while to investigate some of the statistical problems of estimation of linkage disequilibrium.

The degree of linkage disequilibrium can be estimated directly from the genotypic frequencies in a sample of individuals taken from the population. The coupling and repulsion heterozygotes can not normally be distinguished, however, and if either locus is dominant (which for electrophoretic variants usually implies the existence of null alleles) other classes are also confounded.

An alternative approach which is only applicable to *Drosophila* is the isolation of single chromosomes from natural populations against crossover-suppressor stocks. These single chromosomes may thus be made homozygous before establishing their allelic content (*e.g.* Kojima *et al.*, 1970; Mukai *et al.*, 1971). An equivalent procedure is to test cross individuals against a marker stock. The technique of chromosome isolation, in particular, involves much more labour per observation, *i.e.* a diploid or a haploid (chromosome) individual identified, and we may ask whether this labour is justified in terms of improved accuracy of estimation of the disequilibrium. This question was raised with me by Dr D. A. Briscoe, and an attempt is made to provide an answer in this paper by predicting the sampling variance of estimates of disequilibrium obtained by the alternative methods.

It is recommended that maximum likelihood (ML) estimation be used in any such analysis of data, for even where numerical solutions are required these can be obtained easily using relevant computer programs. (A program specifically for handling the analysis of designs discussed in this paper is available from the author.) Whilst the main results of this paper are predictions of sampling variances, it has been extended to include methods of estimation, together with examples to help the experimentalist. For the case of two codominant loci an ML procedure has been given by Bennett (1965), but an alternative method is presented here; and the ML solution for two dominant loci has been given by Turner (1968) and Cavalli-Sforza and Bodmer (1971) but is repeated for completeness.

2. Analysis

The population is assumed to be random mating and to be in Hardy-Weinberg equilibrium at each locus. At the first locus there are two alleles, A and a, with frequencies p and $1-p$, and at the second locus two alleles, B and b, with frequencies q and $1-q$. The frequencies of the chromosome types AB, Ab, aB and ab are f_{11}, f_{12}, f_{21} and f_{22} respectively, and the linkage disequilibrium, D, is given by

$$D = f_{11}f_{22} - f_{12}f_{21} = f_{11} - pq.$$

The frequencies are summarised in table 1 (*a*). We shall alternate between use of the (f_{ij}) and (p, q, D) to define the model, according to which gives the more condensed form of results, and utilise the property that the same transformation applied to the ML estimators $(\hat{f}_{ij})$ gives the ML estimators $(\hat{p}, \hat{q}, \hat{D})$, and vice versa (*e.g.* Elandt-Johnson, 1971, p. 298).

We consider three models in which diploid individuals are identified: both A and B codominant (where the ML estimation procedure is outlined more fully), A codominant and B dominant, and then both A and B dominant. Finally we consider the case where haploids are identified, either by isolation of chromosomes or by appropriate test crossing. In all cases the numbers of each type identified are assumed to be multinomially distributed.

(i) *Diploid identification: both* A *and* B *to dominant*

When all three genotypes can be identified at both loci, but the coupling and repulsion heterozygotes can not be separated, there are nine phenotypic classes. The expected frequencies (y_{ij}, where $y_{11} = f_{11}^2$, for example), the

TABLE 1

Expected frequencies and observed numbers for different genetic models

(a) *Definitions of frequencies; chromosome identification*

Chromosome	*AB*	*Ab*	*aB*	*ab*	Total
Expected frequency	f_{11} $pq+D$	f_{12} $p(1-q)-D$	f_{21} $(1-p)q-D$	f_{22} $(1-p)(1-q)+D$	
Observed numbers	n_{11}	n_{12}	n_{21}	n_{22}	n

(b) A *codominant*, B *codominant: expected frequencies* (γ_{ij})

	BB	*Bb*	*bb*
AA	f_{11}^2	$2f_{11}f_{12}$	f_{12}^2
Aa	$2f_{11}f_{21}$	$2f_{11}f_{22}+2f_{12}f_{21}$	$2f_{12}f_{22}$
aa	f_{21}^2	$2f_{21}f_{22}$	f_{22}^2

(c) A *codominant*, B *codominant: observed numbers*

	BB	*Bb*	*bb*	Total
AA	N_{11}	N_{12}	N_{13}	$N_{1\cdot}$
Aa	N_{21}	N_{22}	N_{23}	$N_{2\cdot}$
aa	N_{31}	N_{32}	N_{33}	$N_{3\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	N

Derived totals

$X_{11} = 2N_{11}+N_{12}+N_{21}$; $X_{12} = 2N_{13}+N_{12}+N_{23}$
$X_{21} = 2N_{31}+N_{21}+N_{31}$; $X_{22} = 2N_{33}+N_{23}+N_{32}$

(d) A *codominant*, B *dominant: observed numbers* (expected frequencies are obtained by summing columns 1 and 2 in (b))

	B–	*bb*	Total
AA	N_{11}	N_{12}	$N_{1\cdot}$
Aa	N_{21}	N_{22}	$N_{2\cdot}$
aa	N_{31}	N_{32}	$N_{3\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	N

(e) A *dominant*, B *dominant: observed numbers* (expected frequencies are obtained by summing rows 1 and 2 and columns 1 and 2 in (b))

	B–	*bb*	Total
A–	N_{11}	N_{12}	$N_{1\cdot}$
aa	N_{21}	N_{22}	$N_{2\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	N

observed numbers (N_{ij}) and some functions of them (X_{ij}) are given in table 1 (*b*) and (*c*). The logarithm of the likelihood (L) is

$$\log L = \sum_{i,j=1}^{3} N_{ij} \log y_{ij} + \text{constant}$$

$$= \sum_{ij} X_{ij} \log f_{ij} + N_{22} \log (f_{11}f_{22}+f_{12}f_{21}) + \text{constant}, \quad (1)$$

which has been given by Bennett (1965). The parameter estimates can be obtained by differentiating log L, and finding the zero values by trial and error, as Bennett (1965) showed. Alternatively we can use the " gene-counting " method of Ceppellini, Siniscalco and Smith (1955) and described

*

by Elandt-Johnson (1971, p. 400), which gives identical solutions to maximum likelihood. Since it is applied in this paper to chromosomes we shall call it the " chromosome counting " method, and it appears to have been used by Webster (1973). Each phenotypic class is apportioned into the expected number of each chromosome type; thus an *AABb* individual comprises one *AB* and one *Ab* chromosome, while *AaBb* individuals have an expected proportion of $f_{11}f_{22}/(f_{11}f_{22}+f_{12}f_{21})$ *AB* and *ab* chromosomes and $f_{12}f_{21}/(f_{11}f_{22}+f_{12}f_{21})$ *Ab* and *aB* chromosomes. The equations are then

$$\begin{aligned}\hat{f}_{ij} &= [X_{ij}+N_{22}\hat{f}_{11}\hat{f}_{22}/(\hat{f}_{11}\hat{f}_{22}+\hat{f}_{12}\hat{f}_{21})]/2N, \quad i=j \\ \hat{f}_{ij} &= [X_{ij}+N_{22}\hat{f}_{12}\hat{f}_{21}/(\hat{f}_{11}\hat{f}_{22}+\hat{f}_{12}\hat{f}_{21})]/2N, \quad i\neq j.\end{aligned} \tag{2}$$

By summing equations (2) we find that the gene frequency estimates are given by the marginal frequencies:

$$\begin{aligned}\hat{p} &= \hat{f}_{11}+\hat{f}_{12} = (X_{11}+X_{12}+N_{22})/2N = (N_{1\cdot}+\tfrac{1}{2}N_{2\cdot})/N, \\ \hat{q} &= \hat{f}_{11}+\hat{f}_{21} = (N_{\cdot 1}+\tfrac{1}{2}N_{\cdot 2})/N;\end{aligned} \tag{3}$$

but $\hat{D}$ has no explicit solution. A suitable method is to replace $\hat{f}_{12}$ by $\hat{p}-\hat{f}_{11}$, $\hat{f}_{21}$ by $\hat{q}-\hat{f}_{11}$ and $\hat{f}_{22}$ by $1-\hat{p}-\hat{q}+\hat{f}_{11}$ in the equation (2) for $\hat{f}_{11}$, to give a single equation

$$\hat{f}_{11} = \{X_{11}+N_{22}\hat{f}_{11}(1-\hat{p}-\hat{q}+\hat{f}_{11})/[\hat{f}_{11}(1-\hat{p}-\hat{q}+\hat{f}_{11})+(\hat{p}-\hat{f}_{11}) \times(\hat{q}-\hat{f}_{11})]\}/2N. \tag{4}$$

The only unknown in (4) is $\hat{f}_{11}$, and it is solved by choosing a value of $\hat{f}_{11}$ for the right-hand side, evaluating the expression and using this as the next trial value of $\hat{f}_{11}$. The iterative process is continued until stability is reached and $\hat{D}$ obtained as $\hat{f}_{11}-\hat{p}\hat{q}$. A suitable starting value for iteration is

$$\hat{f}_{11} = \frac{1}{4N}(X_{11}-X_{12}-X_{21}+X_{22})+\tfrac{1}{2}-(1-\hat{p})(1-\hat{q}), \tag{5}$$

which is obtained by assuming that the genotype frequency of the double heterozygote class is exactly that computed from the other classes.

The sampling variances of the ML estimators can be obtained for large samples in the usual way from the inverse of the matrix of expected values of the log likelihood. Let $t_1 = p$, $t_2 = q$ and $t_3 = D$. From (1)

$$\frac{\partial^2 \log L}{\partial t_k \partial t_l} = \sum_{i,j=1}^{3} N_{ij}\left(y_{ij}\frac{\partial^2 y_{ij}}{\partial t_k \partial t_l} - \frac{\partial y_{ij}}{\partial t_k}\frac{\partial y_{ij}}{\partial t_l}\right)\Big/ y_{ij}^{\,2}$$

We have $E(N_{ij}) = Ny_{ij}$, and note that $\sum_{i,j} \partial^2 y_{ij}/\partial t_k \partial t_l = 0$, since $\sum_{i,j} y_{ij} = 1$ (Elandt-Johnson, 1971, p. 317). Letting

$$m_{kl} = -E(\partial^2 \log L/\partial t_k \partial t_l)$$

we obtain

$$m_{kl} = N \sum_{i,j=1}^{3} \frac{\partial y_{ij}}{\partial t_k}\frac{\partial y_{ij}}{\partial t_l}/y_{ij}. \tag{6}$$

The variance-covariance matrix of the estimates is given by M^{-1}, where M is a 3×3 matrix with elements m_{kl}. The necessary derivatives, $\partial y_{ij}/\partial t_k$, are given in table 2, and these can be used in (6).

TABLE 2

Derivatives of genotypic frequencies (y_{ij}) *for diploid model with both loci codominant with respect to the frequency of* A(p), B(q) *and* D

	BB	*Bb*	*bb*
		$\frac{1}{2}\partial y_{ij}/\partial p$	
AA	qf_{11}	$qf_{12}+(1-q)f_{11}$	$(1-q)f_{12}$
Aa	$q(f_{21}-f_{11})$	$q(f_{22}-f_{12})+(1-q)(f_{21}-f_{11})$	$(1-q)(f_{22}-f_{12})$
aa	$-qf_{21}$	$-qf_{22}-(1-q)f_{21}$	$-(1-q)f_{22}$
		$\frac{1}{2}\partial y_{ij}/\partial q$	
AA	pf_{11}	$p(f_{12}-f_{11})$	$-pf_{12}$
Aa	$pf_{21}+(1-p)f_{11}$	$p(f_{22}-f_{21})+(1-p)(f_{12}-f_{11})$	$-pf_{22}-(1-p)f_{12}$
aa	$(1-p)f_{21}$	$(1-p)(f_{22}-f_{21})$	$-(1-p)f_{22}$
		$\frac{1}{2}\partial y_{ij}/\partial D$	
AA	f_{11}	$f_{12}-f_{11}$	$-f_{12}$
Aa	$f_{21}-f_{11}$	$f_{11}-f_{12}-f_{21}+f_{22}$	$f_{12}-f_{22}$
aa	$-f_{21}$	$f_{21}-f_{22}$	f_{22}

The above method for finding the variances and covariances provides a simple way of computing $V(\hat{D})$ in this codominant-codominant model, and is useful in the other models for parameters which do not have explicit ML estimators. However, for those that do, a direct approach can be used; for example $\hat{p}$ is given by (3) and is binomially distributed. We obtain

$$V(\hat{p}) = p(1-p)/2N, \quad V(\hat{q}) = q(1-q)/2N \tag{7}$$

$$\text{cov}(\hat{p}, \hat{q}) = D/2N, \quad \text{cov}(\hat{p}, \hat{D}) = (1-2p)D/2N, \quad \text{cov}(\hat{q}, \hat{D}) = (1-2q)D/2N.$$

The variances in (7) are, of course, the same as for a single gene situation. When $D = 0$, we see that the covariances are zero, and also find that the equation (6) simplifies, to give

$$V(\hat{D}) = p(1-p)q(1-q)/N. \tag{8}$$

More generally, for $D \neq 0$ it is clear that $V(\hat{D})$ can not be expressed as a linear function of the terms obtained subsequently in (22) for the haploid model.

In any experiment only estimates of p, q and D are available, and these have to be used instead of the parameters in table 2, (6) and (7). Alternatively the second derivatives of the log likelihood can be obtained numerically and used as the elements of M.

Using the large sample assumption of normality, a test for $D = 0$ can be made using (8). This is equivalent to the likelihood ratio test for, under the null hypothesis that $D = 0$, the quantity given by

$$k = -2 \log [L(p, q, D)/L(p, q)] \tag{9}$$

has the chi-square distribution asymptotically with 1 d.f., where $L(p, q, D)$,

$L(p, q)$ are the likelihoods (1) obtained by fitting only the specified parameters. It can be shown that, ignoring terms of order D^3 or higher,

$$k = N\hat{D}^2/\hat{p}(1-\hat{p})\hat{q}(1-\hat{q})$$
$$= N\hat{r}^2, \tag{10}$$

where r^2 is the squared correlation of gene frequencies. The chi-square test proposed by Sinnock and Sing (1972*b*) is equivalent except theirs is obtained by using goodness-of-fit rather than likelihood arguments.

(ii) *Diploid identification:* A *codominant*, B *dominant*

There are now six phenotypes, with the observed numbers shown in table 1 (d) and expected frequencies obtained by summing the appropriate frequencies for B codominant in table 1 (b) (*i.e.* columns 1 and 2). The likelihood equation can be written down using these frequencies but, for solving the equation, we again adopt the chromosome counting method. The equations are (ignoring " hats " on estimates)

$$f_{11} = \frac{1}{2N}\left[\frac{2N_{11}(f_{11}^2+f_{11}f_{12})}{f_{11}^2+2f_{11}f_{12}} + \frac{N_{21}(f_{11}f_{21}+f_{11}f_{22})}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}}\right] \tag{11a}$$

$$f_{12} = \frac{1}{2N}\left[\frac{2N_{11}f_{11}f_{12}}{f_{11}^2+2f_{11}f_{12}} + 2N_{12} + \frac{N_{21}f_{12}f_{21}}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}} + N_{22}\right] \tag{11b}$$

$$f_{21} = \frac{1}{2N}\left[\frac{N_{21}(f_{11}f_{21}+f_{12}f_{21})}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}} + \frac{2N_{31}(f_{21}^2+f_{21}f_{22})}{f_{21}^2+2f_{21}f_{22}}\right] \tag{11c}$$

$$f_{22} = \frac{1}{2N}\left[\frac{N_{21}f_{11}f_{22}}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}} + N_{22} + \frac{2N_{31}f_{21}f_{22}}{f_{21}^2+2f_{21}f_{22}} + 2N_{32}\right]. \tag{11d}$$

Summing equations (11a) and (11b), we find that for the codominant gene, A, the estimated frequency, $\hat{p}$, is given by the marginal frequencies,

$$\hat{p} = (N_{1.} + \tfrac{1}{2}N_{2.})/N. \tag{12}$$

But we notice that the sum of (11a) and (11c) does not simplify in this way, so we obtain the rather surprising result that the ML estimator of gene frequency of a dominant gene suspected of being in disequilibrium with a codominant gene is not given by the marginal frequencies. Similarly, $\hat{D}$ is not obtained explicitly, so we need to retain two of the equations (11), for example (11a) and (11c) and express $\hat{f}_{12}$ and $\hat{f}_{22}$ in terms of $\hat{p}$, $\hat{f}_{11}$ and $\hat{f}_{21}$. These equations are iterated to obtain a solution for $\hat{f}_{11}$ and $\hat{f}_{21}$ and consequently $\hat{q}$ and $\hat{D}$. Since $\hat{q}$ is unlikely to depart far from the estimate given by the marginal frequencies, a suitable starting value for the iterations is obtained using $1-\hat{q} = (N_{.2}/N)^{\frac{1}{2}}$ and $\hat{f}_{22} = (N_{32}/N)^{\frac{1}{2}}$.

The sampling variances of all of the estimators can be found as before, using (6), but with the subscript j taking only two values. The appropriate frequencies y_{ij} and derivatives $\partial y_{ij}/\partial t_k$ are given by summing the first two columns in tables 1 (b) and 2, respectively. Explicit formulae for the variances

or covariances involving the codominant gene A can be given, however. These are the same as when B is codominant also, *i.e.*

$$V(\hat{p}) = p(1-p)/2N \tag{13}$$

$$\text{cov}(\hat{p}, \hat{q}) = D/2N, \quad \text{cov}(\hat{p}, \hat{D}) = (1-2p)D/2N.$$

When $D = 0$, all covariances are zero and

$$V(\hat{q}) = q(2-q)/4N, \quad V(\hat{D}) = p(1-p)q(2-q)/2N; \tag{14}$$

and we note that $V(\hat{q})$ is that for a single dominant gene.

The likelihood ratio criterion (9) for testing $D = 0$ is, approximately,

$$k = 2N\hat{D}^2/[\hat{p}(1-\hat{p})\hat{q}(2-\hat{q})]. \tag{15}$$

(iii) *Diploid identification: both* A *and* B *dominant*

There are only four phenotypic classes (table 1 (e)), so the ML estimators are the obvious ones, namely

$$\hat{p} = 1-(N_{2.}/N)^{\frac{1}{2}}, \quad \hat{q} = 1-(N_{.2}/N)^{\frac{1}{2}} \quad \text{and} \quad \hat{f}_{22} = (N_{22}/N)^{\frac{1}{2}} \tag{16}$$

giving

$$\hat{D} = (N_{22}/N)^{\frac{1}{2}} - (N_{2.}N_{.2})^{\frac{1}{2}}/N \tag{17}$$

(Turner, 1968; Cavalli-Sforza and Bodmer, 1971).

The sampling variances of the estimators can be found using (6), but after summing the first two rows and columns in tables 1 (b) and 2. The only explicit formulae not involving a large number of terms are

$$V(\hat{p}) = p(2-p)/4N, \quad V(\hat{q}) = q(2-q)/4N \tag{18}$$

and the estimators are correlated. When $D = 0$, $\hat{p}$, $\hat{q}$ and $\hat{D}$ are uncorrelated and

$$V(\hat{D}) = p(2-p)q(2-q)/4N. \tag{19}$$

The likelihood ratio criterion (9) is, approximately,

$$k = 4N\hat{D}^2/[\hat{p}(2-\hat{p})\hat{q}(2-\hat{q})], \tag{20}$$

which differs from that given by Cavalli-Sforza and Bodmer (1971, p. 285) in that a term in D^3 has been ignored.

(iv) *Haploid identification*

A sample of n chromosomes is taken from the population and identified by an appropriate method (*e.g.* by test crossing or making an isogenic line) with the observed numbers shown in table 1 (a). The observed chromosome frequencies are their ML estimators, *i.e.* $\hat{f}_{ij} = n_{ij}/n$, so

$$\hat{p} = n_{1.}/n, \quad \hat{q} = n_{.1}/n, \quad \hat{D} = (n_{11}n_{22} - n_{12}n_{21})/n^2. \tag{21}$$

The sampling variances of the estimators can be found directly from the multinomial distribution, with that for $V(\hat{D})$ being obtained from formulae given by Hill and Robertson (1968):

$$\left.\begin{aligned} &V(\hat{p}) = p(1-p)/n, \quad V(\hat{q}) = q(1-q)/n \\ &V(\hat{D}) = [p(1-p)q(1-q) + (1-2p)(1-2q)D - D^2]/n \\ &\text{cov}(\hat{p}, \hat{q}) = D/n, \quad \text{cov}(\hat{p}, \hat{D}) = (1-2p)D/n, \quad \text{cov}(\hat{q}, \hat{D}) = (1-2q)D/n \end{aligned}\right\} \tag{22}$$

We note that, when $D = 0$, the estimates are uncorrelated and

$$V(\hat{D}) = p(1-p)q(1-q)/n. \tag{23}$$

The likelihood ratio criterion (9) is, approximately,

$$k = n\hat{r}^2$$

and k is the usual chi-square statistic in a 2×2 contingency table (Hill and Robertson, 1968).

3. Example

Suitable data for diploid models have been given by Cleghorn (1960) on the *M/N*, *S/s* blood systems in man, and these were also used by Bennett (1965). The data are given in table 3 (a), and we note that both loci are codominant.

Table 3(a)

Cleghorn's data on numbers observed for the M/N *and* S/s *loci and the designation of the alleles in this paper*

Genotype	Designation	*SS* *BB*	*Ss* *Bb*	*ss* *bb*	Total
MM	*AA*	57	140	101	298
MN	*Aa*	39	224	226	489
NN	*aa*	3	54	156	213
	Total	99	418	483	1000

$X_{11} = 293$ $X_{12} = 568$ $X_{21} = 99$ $X_{22} = 592$

Data in 3(a) *reallocated:*

3(b) B *dominant*

	B–	*bb*	Total
AA	197	101	298
Aa	263	226	489
aa	57	156	213
Total	517	483	1000

3(c) A *and* B *dominant*

	B–	*bb*	Total
A–	460	327	787
aa	57	156	213
Total	517	483	1000

(i) A *and* B *codominant*

From (3), $\hat{p} = 0{\cdot}5425$ and $\hat{q} = 0{\cdot}3080$, and with these values inserted into (4) we obtain the chromosome counting formula for iteration

$$\hat{f}_{11} = 0{\cdot}1465 + 0{\cdot}112\hat{f}_{11}(0{\cdot}1495 + \hat{f}_{11})/(0{\cdot}16709 - 0{\cdot}701\hat{f}_{11} + 2\hat{f}_{11}^2)$$

The starting value (5) is $\hat{f}_{11} = 0{\cdot}23791$. After 11 iterations successive values of $\hat{f}_{11}$ differed by less than 10^{-8}, giving a solution of $\hat{f}_{11} = 0{\cdot}2370976$; and from that $\hat{D} = 0{\cdot}0700076$, agreeing with Bennett's value of $\hat{D} = 0{\cdot}07001$. The estimates, together with their standard errors and correlations (computed by replacing the parameter values by their estimates in (6), or in (7) where possible), are summarised in table 4. More figures than are significant are shown for comparison with estimates from the other models. We see in table 4 that D differs significantly ($P < 0{\cdot}001$) from zero, using the likelihood ratio (9) or the approximation to it (10). As Bennett (1965) showed with this data, there is a good fit to Hardy-Weinberg equilibrium: the residual chi-square (from likelihood ratio test) after fitting p, q and D is 3·3 with

5 d.f.). Bennett (1965) gave the standard error of $\hat{D}$ as 0·00596; this value differs slightly from that in table 4, largely because Bennett ignored covariances between the estimators: he assumed $V(\hat{D}) = m_{33}^{-1}$, which he computed by differentiating the likelihood directly.

TABLE 4

Results of analysis of data of table 3

Loci codominant		*A*, *B*	*A*	—
dominant		—	*B*	*A*, *B*
Estimates	p	0·54250	0·54250	0·53848
	q	0·30800	0·30474	0·30502
	D	0·07001	0·07048	0·07422
Standard errors	p	0·01114	0·01114	0·01403
	q	0·01032	0·01135	0·01137
	D	0·00617	0·00712	0·00763
Correlations	p, q	0·3044	0·2788	0·2596
	p, D	−0·0433	−0·0378	−0·1170
	q, D	0·2111	0·1656	0·1725
$-2\ log\ [L(p, q, D)/L(p, q)]$		101·9	79·7	69·3
k (equation 20)		92·6	77·5	54·2

(ii) A *codominant*, B *dominant*

We assume *BB* and *Bb* can not be distinguished in the data in table 3 (a), so by summing the first and second columns we obtain table 3 (b). For gene *A*, $\hat{p} = 0{\cdot}5425$ as before (12). Using (11a) and (11c) and writing $\hat{f}_{12} = \hat{p} - \hat{f}_{11}, \hat{f}_{22} = 1 - \hat{p} - \hat{f}_{21}$ we have

$$\hat{f}_{11} = \frac{0{\cdot}106872}{1{\cdot}0850 - \hat{f}_{11}} + \frac{0{\cdot}060161\hat{f}_{11}}{K}, \quad \hat{f}_{21} = \frac{0{\cdot}026077}{0{\cdot}9150 - \hat{f}_{21}} + \frac{0{\cdot}071338\hat{f}_{21}}{K},$$

where $K = 0{\cdot}4575\,\hat{f}_{11} + 0{\cdot}5425\,\hat{f}_{21} - \hat{f}_{11}\hat{f}_{21}$. Suitable starting values for the iterations are $q = 1 - \sqrt{(483/1000)} = 0{\cdot}3050$ from the marginal totals and $f_{22} = \sqrt{(156/1000)} = 0{\cdot}3950$, equivalent to $\hat{D} = 0{\cdot}0770$, $\hat{f}_{11} = 0{\cdot}2425$, $\hat{f}_{21} = 0{\cdot}0625$. After 22 iterations both $\hat{f}_{11}$ and $\hat{f}_{21}$ changed by less than 10^{-8} in successive iterations, giving, as final values $\hat{q} = 0{\cdot}30474$ and $\hat{D} = 0{\cdot}07048$ (table 4). Notice that the ML estimate of q departs slightly from that computed from marginal frequencies. The data still show a highly significant departure from linkage equilibrium.

(iii) *Both* A *and* B *dominant*

Further reduction of table 3 (a) gives the necessary data for the example in table 3 (c). The ML estimates from (16) and (17) and their sampling variances are listed in table 4. The departure from linkage equilibrium is shown to be significant.

Since the computations are so simple, no example for the chromosomal analysis will be given.

4. DISCUSSION AND CONCLUSIONS

The main object of this analysis was to compare the relative efficiencies of the alternative methods of estimating D. Formally, we measure efficiency as $E = [V(\hat{D})$ from n haploids$]/[V(\hat{D})$ from N diploids$]$, so that $E > 1$ if

the diploid method gives a lower variance for the same number of observations, and $E<1$ if the haploid method gives a lower variance. We recall that a single observation is either the identification of one diploid individual, or the identification of the allelic content of one chromosome, which may be one observation on an isogenic line or one test cross progeny.

The case of most interest is where the population is near linkage equilibrium, or we wish to test the null hypothesis that $D = 0$, and fortunately this has given us the simplest solutions. The results can be summarised as follows:

Haploid identification:

$$V(\hat{D}) = p(1-p)q(1-q)/n = nV(\hat{p})V(\hat{q}).$$

Diploid identification:

$$V(\hat{D}) = 4NV(\hat{p})V(\hat{q})$$

and the efficiencies for the different models are related to the accuracy of gene frequency estimation:

A, B codominant	$E = 1$
A codominant, B dominant	$E = (1-q)/(1-\frac{1}{2}q)$
A, B dominant	$E = [(1-p)/(1-\frac{1}{2}p)][1-q)/(1-\frac{1}{2}q)]$

If both loci are codominant, typical for biochemical variants, we see that $\hat{D}$ has the same variance when estimated from diploids directly as from a sample of the same size of extracted chromosomes or test crosses, which requires much more labour. Some examples have also been computed for $D \neq 0$ for the double codominant case, with p, q = 0·1, 0·25, 0·5 and $q<p$. It turns out that $E \leqq 2$, only approaching $E = 2$ with $p = q = 0{\cdot}5$ and $D \to \pm 0{\cdot}25$, but $E>1$ over most combinations of p, q and D. The only cases with $E<1$ are listed below, together with the lowest values attained:

$(p, q) = (0{\cdot}1, 0{\cdot}1)$, $-0{\cdot}010 < D < 0$, minimum $E = 0{\cdot}74$
$(p, q) = (0{\cdot}25, 0{\cdot}1)$, $-0{\cdot}018 < D < 0$, minimum $E = 0{\cdot}91$
$(p, q) = (0{\cdot}25, 0{\cdot}25)$, $-0{\cdot}031 < D < 0$, minimum $E = 0{\cdot}97$.

Therefore, even when $D \neq 0$, the diploid method is likely to give better estimates, $\hat{D}$, for a given input of labour.

Returning to the case of $D = 0$ and considering dominant genes, we see that the diploid and haploid models have similar efficiencies if the dominant genes are at low frequency; but if they are at high frequency, the chromosome or test cross method may be worth while, just as it would be if we were interested in estimating gene frequencies.

This analysis has been restricted to two loci, but some preliminary studies have been carried out with more. It appears that, if all loci are codominant, the efficiency of the diploid relative to haploid method of estimating the disequilibrium between c loci, under the null hypothesis of equilibrium, is equal to 2^{2-c}. This equals 1 for 2 loci, $\frac{1}{2}$ for 3 loci, $\frac{1}{4}$ for 4 loci, and so on. Thus for three loci the haploid method would be justified only if it required less than twice the labour, per individual scored, than the diploid method. It is interesting to note that the diploid method is twice as efficient for estimating gene frequencies, since two genes are scored per individual, and this efficiency of 2 is obtained by setting $c = 1$ in the above formula. In effect we lose half

the information on D in the two locus diploid cases because we cannot distinguish between the coupling and repulsion heterozygotes, and a greater proportion with more loci when there are several multiple heterozygote classes.

5. References

ALLARD, R. W., BABBEL, G. R., CLEGG, M. T, AND KAHLER, A. L. 1972. Evidence for co-adaptation in *Avena barbata*. *Proc. Nat. Acad. Sci. U.S.A.*, *69*, 3043-3048.

BENNETT, J. H. 1965. Estimation of the frequencies of linked gene pairs in random mating populations. *Amer. J. Hum. Genet.*, *17*, 51-53.

CAVALLI-SFORZA, L. L., AND BODMER, W. F. *The Genetics of Human Populations*. Freeman, San Francisco.

CEPPELLINI, R., SINISCALCO, M., AND SMITH, C. A. B. 1955. The estimation of gene frequencies in a random-mating population. *Ann. Eugen.*, *20*, 97-115.

CHARLESWORTH, B., AND CHARLESWORTH, D. 1973. A study of linkage disequilibrium in populations of *Drosophila melanogaster*. *Genetics*, *73*, 351-359.

CLEGHORN, T. E. 1960. MNSs gene frequencies in English blood donors. *Nature*, *187*, 701.

ELANDT-JOHNSON, R. C. 1971. *Probability Models and Statistical Methods in Genetics*. Wiley, New York.

FRANKLIN, I. R. 1973. Selection, migration and genetic drift in natural populations of *D. melanogaster*. *Genetics*, *74*, s84.

HILL, W. G., AND ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, *38*, 226-231.

KOJIMA, K., GILLESPIE, J., AND TOBARI, Y. N. 1970. A profile of *Drosophila* species' enzymes assayed by electrophoresis. I. Number of alleles, heterozygosities and linkage disequilibrium in glucose-metabolising systems and some other enzymes. *Biochem. Genet.*, *4*, 626-637.

MUKAI, T., METTLER, L. E., AND CHIGUSA, S. I. 1971. Linkage disequilibrium in a local population of *Drosophila melanogaster*. *Proc. Nat. Acad. Sci. U.S.A.*, *69*, 2474-2478.

PRAKASH, S., AND LEWONTIN, R. C. 1968. A molecular approach to the study of genetic heterozygosity in natural populations. III. Direct evidence of coadaptation in gene arrangements of *Drosophila*. *Proc. Nat. Acad. Sci. U.S.A.*, *59*, 398-405.

PRAKASH, S., AND LEWONTIN, R. C. 1971. A molecular approach to the study of genetic heterozygosity in natural populations. V. Further direct evidence of coadaptation in inversions of *Drosophila*. *Genetics*, *69*, 405-408.

SINNOCK, P., AND SING, C. F. 1972*a*. Analysis of multilocus genetic systems in Tecumseh, Michigan. I. Definition of the data set and tests for goodness-to-fit to expectations based on gene, gamete and single locus phenotypic frequencies. *Amer. J. Hum. Genet.*, *24*, 381-392.

SINNOCK, P., AND SING, C. F. 1972*b*. Analysis of multilocus genetic systems in Tecumseh, Michigan. II. Consideration of the correlation between non-alleles in gametes. *Amer. J. Hum. Genet.*, *24*, 393-415.

TURNER, J. R. G. 1968. On supergenes. II. The estimation of gametic excess in natural populations. *Genetica*, *39*, 82-93.

WEBSTER, T. P. 1973. Adaptive linkage disequilibrium between two esterase loci of a salamander. *Proc. Nat. Acad. Sci. U.S.A.*, *70*, 1156-1160.

ZOUROS, E., AND KRIMBAS, C. B. 1972. Linkage disequilibrium in natural populations of *Drosophila subobscura* maintained by selection. *Genetics*, *71*, s71.

PRINTED IN GREAT BRITAIN BY T. & A. CONSTABLE LTD., EDINBURGH

30

Tests for association of gene frequencies at several loci in random mating diploid populations

by

William G. Hill

881-888
Biometrics 31, 000-000
December 1975

TESTS FOR ASSOCIATION OF GENE FREQUENCIES AT SEVERAL LOCI IN RANDOM MATING DIPLOID POPULATIONS

William G. Hill

Institute of Animal Genetics, Edinburgh EH9 3JN, Scotland

SUMMARY

Methods are outlined for analyzing data on genotype frequencies at several codominant loci in random mating diploid populations. Maximum likelihood (ML) methods are given for estimating chromosomal frequencies. Using these, a succession of models of assumed independence of gene frequency are fitted. These are based on those used in multi-dimensional contingency tables, and tests for association (linkage disequilibrium), made using likelihood ratios. The methods are illustrated with an example.

1. INTRODUCTION

The technique of gel electrophoresis enables individual animals or plants to be typed for several different polymorphic enzyme loci. Populations can thus be examined to determine whether the frequencies of genes at different loci are independent. In the absence of the disturbing forces of mutation, migration, selection, drift and non-random mating, the frequency f_{AB} of chromosomes carrying genes A and B at a pair of loci approaches the product $f_A f_B$, of the marginal frequencies, and similarly for more loci. A non-zero value of $f_{AB} - f_A f_B$ is usually termed linkage disequilibrium. (For further details see e.g. Crow and Kimura [1970]). From the presence of linkage disequilibrium in different populations it may be possible to deduce information on selection, drift and migration. For example, disequilibrium caused solely by drift would not be expected to be of the same sign or magnitude in different populations, in contrast to that caused by selection. Lewontin [1974] argues that such tests for gene association should be a very powerful method for distinguishing between alternative evolutionary models; Charlesworth and Charlesworth [1973] and Mitton *et al.* [1973], for example, discuss this approach in more detail.

There is a large literature on tests of association in multi-way contingency tables. Lancaster [1951] demonstrated a straightforward partition of the total chi-square, but this was later shown to lead to incorrect tests for second-order (i.e. three-variable association), the correct version initially being given by Bartlett [1935]. Useful expositions are by Goodman [1969] and Fienberg [1970]. In genetic techniques where individual chromosomes are made homozygous, using cross-over suppressor stocks in *Drosophila melanogaster* or in naturally self fertilizing plants, these standard analyses for multi-way contingency tables can be used for testing for disequilibrium (Smouse [1974] and Morgan and Somerville [1974]).

In most species, however, chromosome frequencies can not be estimated directly. The genes determining electrophoretic variants are usually codominant, so that at a locus with s alleles a total of $s(s + 1)/2$ genotypes can be identified. It is not possible to determine whether heterozygotes at two or more loci are in coupling or replusion phase. ML methods,

however, can be used to estimate the chromosome frequencies and test for the presence of linkage disequilibrium. Such procedures have been given for two loci (Bennett [1965]; Hill [1974]), and it has been shown that, using ML, equal information can be extracted about two-locus disequilibrium from a sample of diploids as from a sample of individual chromosomes of the same size, but with the diploid typing requiring much less laboratory work (Hill [1974]).

In this note the methods of ML estimation and testing are extended to more than two loci. The detailed analysis is restricted to three loci each with two alleles for most of the conceptual problems arise at this level. Throughout, the population is assumed to be random mating, with the expected frequencies of diploid individuals equal to the product of their constituent chromosomes (a multi-locus equivalent of Hardy–Weinberg equilibrium). For this model to hold, say in adults, there must have been no migration into the population during their lifetime, nor any substantial selection on viability, which would distort genotype frequencies. As part of the procedure, however, a test for Hardy–Weinberg equilibrium can be made.

2. METHOD

Let α, β and γ be three loci, each with two codominant alleles A/a, B/b and C/c, respectively. A sample of N individuals are typed from a population in which mating is assumed to be random, and of these $n(AaBbCc)$ have genotype $AaBbCc$. The numbers of each genotype are assumed to be multinomially distributed. If the frequency of chromosome type ABC, for example, is f_{ABC}, the expected genotype frequency is

$$g(AaBbCc) = 2(f_{ABC}f_{abc} + f_{ABc}f_{abC} + f_{AbC}f_{aBc} + f_{Abc}f_{aBC}). \tag{1}$$

The expected frequency of ABC chromosomes in $AaBbCc$ individuals is thus

$$h(AaBbCc, ABC) = 2f_{ABC}f_{abc}{}^{-1}g(AaBbCc) = h(AaBbCc, abc). \tag{2}$$

Frequency estimation and computation of likelihoods

The ML solution for the chromosome frequencies can be obtained by chromosome counting, an extension of the gene counting method of Ceppellini *et al.* [1955], in which the frequencies are equated to their expectations in successive approximations until convergence is reached. Letting $\hat{f}_{ABC}$ denote the next approximation to the chromosome frequencies and f_{ABC} the most recent, then $\hat{f}_{ABC}$ is obtained from f_{ABC} according to the relation

$$\hat{f}_{ABC} = \sum_x n(x)h(x, ABC)/2N \tag{3}$$

where x denotes a genotype, for example $AaBbCc$ with $h(x, ABC)$ as given by (2) and $g(x)$ by (1). Summation in (3) is over all genotypes.

In this model chromosome frequencies for all loci have been fitted. The log likelihood is, ignoring constant terms,

$$L(\alpha\beta\gamma) = \sum_x n(x) \ln g(x) \tag{4}$$

where $g(x)$ is computed from (1) using the ML estimates of chromosome frequencies.

In subsequent models, ML estimates are used of gene frequencies and frequencies of chromosomes at which only two loci are identified. These can be obtained using the same

chromosome counting technique, and thus computing routine, on marginal totals. For example, assume only α and β are included. Then we have, from (1),

$$g(AaBb) = 2(f_{AB}f_{ab} + f_{Ab}f_{aB}).$$

The equivalent function to h is obtained by analogy to (2), and marginal totals, e.g. $n(AaBb, .) = n(AaBbCC) + n(AaBbCc) + n(AaBbcc)$, are used in (3). Such estimates of chromosome frequency, e.g. $\hat{f}_{AB.}$, do not in general equal the appropriate marginal totals $\hat{f}_{AB.} = \hat{f}_{ABC} + \hat{f}_{ABc}$ obtained from fitting frequencies of all three loci. Gene frequencies can be obtained by a similar condensation of the chromosome counting technique or explicitly: for example,

$$\hat{f}_A = [n(AA, ., .) + \tfrac{1}{2}n(Aa, ., .)]/N.$$

It is a consequence of the counting procedure that the gene frequencies equal marginal totals such as $\hat{f}_{A.} = \hat{f}_{AB} + \hat{f}_{Ab}$ and $\hat{f}_{A..} = \hat{f}_{ABC} + \hat{f}_{ABc} + \hat{f}_{AbC} + \hat{f}_{Abc}$. The likelihoods on the marginal totals are given by substituting these into (4), and are denoted $L(\alpha)$, $L(\alpha\beta)$, etc.

Models of association of gene frequencies

Successive models of dependence among the gene frequencies can be fitted, but there is no obvious uniquely preferable way of doing this, for the types of association found will depend on the selective and other forces acting on the population and the degree of recombination between the loci. Given a prior hypothesis about any particular set of data specific, appropriate, models could be tested. In the absence of such a hypothesis, or in the general case discussed here, it seems preferable to follow the hierarchy of models of dependence of frequencies considered appropriate for the standard multi-dimensional contingency table (Goodman [1969], Fienberg [1970] and utilized by Smouse [1974] and Morgan and Somerville [1974] for analyzing individual chromosome, as opposed to diploid data). If significant departures from random association are found at any level in the models, the possible selective forces, for example, which could give rise to this can then be studied. We shall give no more than indications of what these might be.

A succession of models are given in Table 1. These follow Fienberg [1970], although we have not used his log-linear arguments. In *model 0* there is complete independence of the genes, so chromosome frequencies equal products of their constituent gene frequencies. This could occur if the three loci were unlinked. In *model 1* there is association between genes at a pair of loci, say α and β, but not at the third, (this is one of three alternative combinations); and a two-way contingency table having 4 rows specifying AB, Ab, aB and ab and two columns C and c with entries equal to the chromosome frequencies would show random association. *Model 1* would be relevant if α and β were linked and interacted due to selection, and γ was unlinked. In *Model 2* there is association between genes at two pairs of loci, say between α and β and between α and γ, but not between the third pair, β and γ (this is also one of three alternative combinations). *Model 2* implies independence of genes at β and γ for a given gene at α, so in a two-way table with rows B, b and columns C, c there would be independence among the four chromosome frequencies f_{ABC}, f_{ABc}, f_{AbC} and f_{Abc} carrying A, and similarly in the table of frequencies of chromosomes having a. *Model 2* could be appropriate if there were no epistasis between genes at β and γ within A or a, but with equilibrium frequencies differing in the two groups; alternatively the population might derive from a recent cross between two populations, one fixed for A the other for a, and with all genes neutral. In *model 3* all pair-wise associations are present but not the three-way.

TABLE 1

SUCCESSION OF MODELS OF ASSOCIATION OF GENE FREQUENCIES

Model	f_{ABC}	$g(AaBbCc)$	Interactions	Likelihood
0	Complete independence of frequencies at α, β and γ			
	$f_A f_B f_C$	$8 f_A f_a f_B f_b f_C f_c$	-	$L_0 = L(\alpha)+L(\beta)+L(\gamma)$
1	Frequencies at γ independent of those at α and β [o]			
	$f_{AB} f_C$	$4 f_C f_c (f_{AB} f_{ab} + f_{Ab} f_{aB})$	αβ	$L_1(\alpha\beta) = L(\alpha\beta)+L(\gamma)$
2	Independence of frequencies at β and γ conditional on frequency at α [o]			
	$f_{AB} f_{AC} / f_A$	$2(f_{AB} f_{ab} + f_{Ab} f_{aB}) \times (f_{AC} f_{ac} + f_{Ac} f_{aC}) / (f_A f_a)$	αβ+αγ	$L_2(\alpha\beta,\alpha\gamma) = L(\alpha\beta) + L(\alpha\gamma) - L(\alpha)$
3	No three-locus associations			
	f'_{ABC} ≠	f'_{ABC} in eq. (1)	αβ+αγ+βγ	L_3
4	All two- and three-locus associations			
	f_{ABC}	eq. (1)	αβ+αγ+βγ+αβγ	$L_4 = L(\alpha\beta\gamma)$

o One of three alternatives

≠ No explicit formula, see eq. (5).

This implies that the chromosome frequencies satisfy

$$f_{ABC}' f_{Abc}' f_{aBc}' f_{abC}' = f_{ABc}' f_{AbC}' f_{aBC}' f_{abc}' \qquad (5)$$

(Bartlett [1935]), but no explicit formula for them can be given. Also there is disequilibrium between each of the marginal pairs of chromosome frequencies. Such a situation might occur in genetic models in which there is two-locus but no three-locus epistasis. In *model 4* all associations, including those relating to three loci, are possible.

In *models 0, 1 and 2* it is seen that the expression for $g(AaBbCc)$ in (1) factorizes. Hence the gene frequencies or two locus chromosome frequencies are estimated using the appropriate marginal totals of the observations, and the log likelihoods are obtained as sums of those on the margins. The problem is to fit model 3, with all two-locus yet no three-locus associations. This satisfies (5) but as there is no explicit formula for the chromosome frequencies, in the standard three dimensional contingency table an iterative technique has to be used to compute these three locus frequencies, with each of the two locus marginal frequencies satisfied. A method is given by Goodman [1969] and Fienberg [1970]. The procedure has to be modified for the diploid situation: with the pair-wise frequencies f_{AB}, f_{AC}, f_{BC} etc. computed for *model 1*, find the frequencies, e.g. f_{ABC}', satisfying *model 3* using the iterative technique given by Fienberg. With these compute L_3 from (4). It has been found that this value is already near the maximum, but further increases in L_3 are achieved by modifying the pairwise frequencies to f_{AB}', f_{AC}', f_{BC}', with the restriction that

the gene frequencies are not changed, and repeating the procedure. A standard computer program for maximizing non-linear functions can be used.

The alternative models can be compared using the likelihood ratio test, assuming that differences in doubled log likelihoods are distributed as chi-square under the appropriate null hypotheses. The statistic for testing association of α and β, assuming independence of both from γ, is thus

$$2[L_1(\alpha\beta) - L_0] = 2[L(\alpha\beta) - L(\alpha) - L(\beta)]$$

which has 1 D.F. A test for independence of α and γ, assuming α and β are associated is given by

$$2[L_2(\alpha\beta, \alpha\gamma) - L_1(\alpha\beta)] = 2[L(\alpha\gamma) - L(\alpha) - L(\gamma)]$$

which is the same as if a possible association of α and β were ignored. The test statistic for independence of β and γ, assuming other pair-wise associations is $2[L_3 - L_2(\alpha\beta, \alpha\gamma)]$. These associations of pairs of loci could be fitted in a different order. The statistic for testing whether there are any three-locus associations is $2(L_4 - L_3)$, again with 1 D.F. (This is different from the statistic given by Lancaster [1951], using analysis of variance arguments, which is $L(\alpha\beta\gamma) - L(\alpha\beta) - L(\alpha\gamma) - L(\beta\gamma) + L(\alpha) + L(\beta) + L(\gamma)$).

3. EXTENSIONS

Hardy–Weinberg association

A final test which may be useful is for a fit to Hardy–Weinberg (H–W) associations of chromosome frequencies, i.e. that genotype frequencies equal the product of chromosome frequencies. If this assumption is removed at *model 4*, in which gene frequencies may also be associated in any way, the log likelihood is

$$L_5 = \sum_x n(x) \ln [n(x)/N]$$

The test statistic is $2(L_5 - L_4)$ with $3^4 - 2^4 = 19$ D.F. If numbers of some of the genotypes are too small for the chi-square test to be appropriate, some pooling of classes may be necessary. Similar tests can also be made using marginal frequencies of genes or pairs of chromosomes, to find the particular loci contributing to H–W disequilibrium if it is found in the above test. Should any significant departure from H–W equilibrium be found in a set of data, the validity of the tests for gene association shown in Table 1 should be questioned, for the gene and chromosome frequencies estimated by ML assume H–W equilibrium.

Multiple alleles

The methodology can readily be extended to include more than two alleles at a locus, for the chromosome counting method and the sequence of models still apply, although the degrees of freedom have to be modified. The computation of L_3 is more lengthy since it has to be maximized with respect to more marginal frequencies.

More than three loci

This extension can also be carried out, and the relevant hierarchy of models for contingency tables is given by Fienberg [1970]. For four loci there are now six likelihoods such

as $L(\alpha\beta)$ and four such as $L(\alpha\beta\gamma)$ and L_3. In addition there are new quantities in which all six two-locus disequilibria are fitted, and in which all two- and all three-locus disequilibria are fitted. In principle a combination of the iterative routine of Fienberg together with use of a maximizing routine can be used, but in practice the computations soon become too tedious. It is probable that the best procedure would be just to fit the likelihood containing all pairs, and fit $L(\alpha\beta\gamma\delta)$ where δ is the fourth locus, and compare these, for the difference contains all three and four-locus disequilibria.

4. EXAMPLE

Data for this example were kindly provided by J. B. Mitton. They are on the coastal marine fish *Fundulus heteroclitus* from Northport on Long Island, New York in 1971 (Mitton and Koehn [1975]). The loci are labelled here as α = lactate hehydrogenase, β = serum esterase, γ = transferrin. Alleles A, B, C, respectively, refer to the faster migrating of the alleles at the locus on the gel and a, b, c to the slower allele. The numbers of each type are given in Table 2; there were 319 individuals typed for these three loci.

At locus α, for example, $\hat{f}_A = (30 + 120/2)/319 = 0.28213$ and

$$L(\alpha) = 30 \ln \hat{f}_A^2 + 120 \ln 2\hat{f}_A\hat{f}_a + 169 \ln \hat{f}_a^2 = -296.404.$$

Fitting α and β by chromosome counting, $\hat{f}_{AB} = 0.21475$, for example, and $L(\alpha\beta) = -562.956$. Fitting all three loci by chromosome counting, the ML estimates of the chromosome frequencies are:

$$\hat{f}_{ABC} = 0.18401, \quad \hat{f}_{ABc} = 0.03086, \quad \hat{f}_{AbC} = 0.04975, \quad \hat{f}_{Abc} = 0.01751$$

$$\hat{f}_{aBC} = 0.46194, \quad \hat{f}_{aBc} = 0.11316, \quad \hat{f}_{abC} = 0.10211, \quad \hat{f}_{abc} = 0.04066$$

and $L(\alpha\beta\gamma) = -817.269$. Note that $\hat{f}_{A..} = 0.28213 = \hat{f}_A$, whereas $\hat{f}_{AB.} = 0.21487 \neq \hat{f}_{AB}$. In *model 3*, in which all pair-wise associations are fitted, at the maximum $f_{ABC}' = 0.18248$ and $f_{AB.}' = 0.21475$, and $L_3 = -817.3205$ which is only marginally greater than the value, -817.3210, obtained by using Fienberg's routine on the marginal totals $\hat{f}_{AB.}$, $\hat{f}_{A.C}$, etc. without going through a further maximization. Results are summarized in Table 3.

The total log likelihood ratio statistic for the 4 D.F. between *model 4* (which includes

TABLE 2

NUMBER OF EACH GENOTYPE IN MITTON'S DATA

	CC			Cc			cc			
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa	Total (β)
BB	14	56	72	4	17	35	1	2	7	208
Bb	4	22	29	3	11	13	2	1	3	88
bb	2	6	3	0	4	6	0	1	1	23
Total (γ)		208			93			18		319
Total (α)	30	120	169							

TABLE 3

ANALYSIS OF MITTON'S DATA

a. Summary of likelihoods

L(α)	L(β)	L(αβ)	L(γ)	L(αγ)
-296.404	-268.934	-562.956	-256.725	-552.662
L(βγ)	L_3	$L(\alpha\beta\gamma)=L_4$	L_5	
-521.861	-817.321	-817.269	-805.027	

b. Likelihood ratio tests

Source		D.F.	Test statistic: Formula	Test statistic: value
Association of α and β[a]		1	$2[L(\alpha\beta)-L(\alpha)-L(\beta)]$	0.763
	α and γ	1	$2[L(\alpha\gamma)-L(\alpha)-L(\gamma)]$	0.934
	β and γ	1	$2[L(\beta\gamma)-L(\beta)-L(\gamma)]$	3.597
Association of: after fitting:				
α and β	αγ,βγ	1	$2[L_3-L(\alpha\gamma)-L(\beta\gamma)+L(\gamma)]$	0.954
α and γ	αβ,βγ	1	$2[L_3-L(\alpha\beta)-L(\beta\gamma)+L(\beta)]$	1.125
β and γ	αβ,αγ	1	$2[L_3-L(\alpha\beta)-L(\alpha\gamma)+L(\alpha)]$	3.788
Three-way association		1	$2(L_4-L_3)$	0.103
Total		4	$2[L_4-L(\alpha)-L(\beta)-L(\gamma)]$	5.588
Hardy-Weinberg fit		19	$2(L_5-L_4)$	24.483

[a] or: association of α and β given either α and γ are associated or β and γ are associated.

all associations) and *model 0* (for independence) is 5.588 and non-significant. The only association approaching significance at the 5 percent level is that between genes at loci β and γ, which after fitting $\alpha\beta$ and $\alpha\gamma$ gives a chi-square statistic of 3.788 with 1 D.F. Since there was so little association of frequencies found in these data, the order in which the pairs are fitted makes little difference, but the most appropriate order seems to be to test the questionable pair last. This gives the sequence, with test statistic: $\alpha\beta$ (0.763), $\alpha\gamma$ after $\alpha\beta$ (0.934), $\beta\gamma$ after $\alpha\beta$ and $\alpha\gamma$ (3.788), $\alpha\beta\gamma$ after $\alpha\beta$, $\alpha\gamma$, $\beta\gamma$ (0.103). Finally, the test for H–W equilibrium gives 24.483 with 19 D.F. which indicates no departure and the assumption of random mating in the analysis seems tenable.

ACKNOWLEDGMENTS

I am indebted to Dr. Jeffry B. Mitton for providing the data and making several helpful comments, and to Mrs. Jennifer Smith for programming the analysis on the computer.

TESTS D'ASSOCIATION DES FREQUENCES GENIQUES A DIFFERENTS LOCUS DANS DES POPULATIONS DIPLOIDES AVEC CROISEMENT AU HASARD

RESUME

Des méthodes sont proposées pour l'analyse des données concernant les fréquences génétypiques à différents locus codominants dans des populations diploïdes en random mating. On donne des méthodes

du maximum de vraisemblance pour estimer les fréquences des chromosomes. A partir de ces fréquences, on ajuste une succession de modèles supposant l'indépendance des fréquences géniques. Ils reposent sur ceux utilisés dans les tableaux de contingence multidimensionnels et sur des tests d'association (déséquilibre ou linkage) faits à partir des rapports de vraisemblance. Les méthodes sont illustrées par un exemple.

REFERENCES

Bartlett, M. S. [1935]. Contingency table interactions. *J. R. Statist. Soc. Suppl. 2*, 248–52.

Bennett, J. H. [1965]. Estimation of the frequencies of linked gene pairs in random mating populations. *Amer. J. Hum. Genet. 17*, 51–3.

Charlesworth, B. and Charlesworth, D. [1973]. A study of linkage disequilibrium in populations of *Drosophila melanogaster*. *Genetics 73*, 351–9.

Ceppellini, R., Siniscalco, M. and Smith, C. A. B. [1955]. The estimation of gene frequencies in a random mating population. *Ann. Eugen. 20*, 97–115.

Crow, J. F. and Kimura, M. [1970]. *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Fienberg, S. E. [1970]. The analysis of multidimensional contingency tables. *Ecology 51*, 419–33.

Goodman, L. A. [1969]. On partitioning χ^2 and detecting partial association in three-way contingency tables. *J. R. Statist. Soc. B 31*, 486–98.

Hill, W. G. [1974]. Estimation of linkage disequilibrium in random mating populations. *Heredity 33*, 229–39.

Lancaster, H. O. [1951]. Complex contingency tables treated by the partition of chi-square. *J. R. Statist. Soc. B 13*, 242–9.

Lewontin, R. C. [1975]. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Mitton, J. B. and Koehn, R. K. [1974]. Genetic organisation and adaptive response of allozymes to ecological variables in *Fundulus heteroclitus*. *Genetics 79*, 97–111.

Mitton, J. B., Koehn, R. K. and Prout, T. [1973]. Population genetics of marine pelecypods. III. Epistasis between functionally related isoenzymes of *Mytilus edulis*. *Genetics 75*, 487–96.

Morgan, K. and Someville, C. R. [1974]. Analysis of linkage disequilibrium in populations of *Drosphila melanogaster*: additional analyses by log-linear models. *Genetics* (submitted).

Smouse, P. E. [1974]. Likelihood analysis of recombinational disequilibrium in multiple-locus gametic frequencies. *Genetics 76*, 557–65.

Received July 1974, *Revised November* 1974